Assessing Procedural Competence

Validity Considerations

Debra M. Pugh, MD, MHPE;
Timothy J. Wood, PhD;
John R. Boulet, PhD

Summary Statement: Simulation-based medical education (SBME) offers opportunities for trainees to learn how to perform procedures and to be assessed in a safe environment. However, SBME research studies often lack robust evidence to support the validity of the interpretation of the results obtained from tools used to assess trainees' skills. The purpose of this paper is to describe how a validity framework can be applied when reporting and interpreting the results of a simulation-based assessment of skills related to performing procedures. The authors discuss various sources of validity evidence because they relate to SBME. A case study is presented.
(*Sim Healthcare* 10:288–294, 2015)

Key Words: Assessment, Procedural skills, Validity

Opportunities for medical trainees to perform procedures on real patients have decreased for the past decade because of patient safety concerns,[1] imposed reduced work hours,[2,3] and shorter hospital stays for patients.[4] In answer to this, simulation-based medical education (SBME) has rapidly been incorporated into many undergraduate and postgraduate training programs, as a tool for both the teaching and assessment of procedural competence. If done well, the incorporation of SBME can provide an opportunity for trainees to acquire and practice their skills in a safe environment, as well as allow educators to assess their trainees and provide directive feedback.[5] However, recent reviews examining the use of simulation as an educational intervention in healthcare have highlighted several weaknesses in SBME research, including the lack of evidence to support the reliability and validity of the scores obtained from tools used to assess trainees' skills.[6–11]

Given that novel assessment tools to be used in SBME are continually being developed and published studies of their merits abound, it can be tempting for educators to rush to adopt the "newest thing."[12] Doing so places the educator at risk of using a flawed instrument that may produce results (eg, assessment scores) that are uninterpretable or invalid and, therefore, of little value to teachers and learners. To infer anything about a trainee's competence, it is imperative that the measurement instrument used provides scores that accurately represent his/her abilities. Thus, gathering evidence to support the use of the tool and the appropriateness of any inferences to be made on the basis of the assessment scores is essential.[13]

One of the challenges with assessing competence is that competence is defined by constructs, which are intangible concepts that can only be measured indirectly. Examples of constructs include intelligence or procedural ability. When assessing constructs, one must make inferences using indirect measures. For example, an IQ test may be used as a measure of intelligence and a procedural objective structured clinical examination may be used to assess procedural ability. Although these are not direct measures, one can make inferences about a person's intelligence or procedural ability using scores obtained from these tests. Because constructs cannot be directly measured and interpretations are based on inferences, one must collect data (ie, evidence) that can either support or refute a particular inference concerning the construct being measured. The evidence that is collected is then evaluated in light of interrelated criteria, such as the 5 sources of validity evidence described in the Standards of Educational and Psychological Testing.[14] The collection of data based on the "standards" or other relevant frameworks[15,16] can help support the validity argument; that is, the scores obtained from the assessment accurately and consistently reflect "true" ability and not some other unintended, or construct-irrelevant, factors.

The purpose of this paper is to describe how a modern validity framework could be applied to the assessment of procedural competence using simulation. Table 1 provides a listing of these sources of validity evidence along with a definition and a description of possible factors that could be a threat to each source of evidence. The use of these sources of evidence to support validity arguments will be discussed in further detail later. A case study of how a modern validity framework could be applied when reporting and interpreting the results of an examination used to assess

From the Departments of Medicine (D.M.P.) and Innovation in Medical Education (T.J.W.), University of Ottawa, Ottawa, ON, Canada; and Foundation for Advancement of International Medical Education and Research (J.R.B.), Philadelphia, PA.

Reprints: Debra M. Pugh, MD, MHPE, Division of General Internal Medicine, The Ottawa Hospital General Campus, 501 Smyth Rd, Box 209, Ottawa, ON K1H 8L6, C (e-mail: dpugh@toh.on.ca).
The authors declare no conflict of interest.

Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's Web site (www.simulationinhealthcare.com).
Copyright © 2015 Society for Simulation in Healthcare
DOI: 10.1097/SIH.0000000000000101

Copyright © 2015 by the Society for Simulation in Healthcare. Unauthorized reproduction of this article is prohibited.

TABLE 1. Demonstrating Validity Evidence

Categories of Validity Evidence	Examples of Validity Evidence	Examples of Threats to Validity
Content evidence		
Demonstrating that the construct is adequately represented	Use of a blueprint based on procedures of interest	Inclusion of too few questions or cases
	Demonstration of case writer qualifications (eg, cases developed by those who regularly perform the procedure)	Use of flawed items (eg, cueing)
	Demonstration of high-quality cases or scenarios (eg, by having a review process)	
	Documentation of a rationale for choice of scoring rubric (eg, by using previously studied tools)	
Response process evidence		
Ensuring integrity of data collection	Use of instructions and prompts that have been reviewed to ensure clarity	Use of unfamiliar equipment and formats
	Use of rater training and evidence of effectiveness of that training (eg, by demonstrating high interrater reliability)	Lack of test security (eg, examinees cheating)
	Use of quality assurance measures (eg, double-checking data entry)	
Internal structure evidence		
Analyzing the psychometric characteristics of the scores obtained	Demonstration of internal consistency (eg, high Cronbach alpha)	Systematic bias against a group (eg, as identified through differential item functioning analysis)
	Analysis of various sources of error (eg, generalizability analysis)	Low reliability
	Demonstration of association between cases and total score (eg, item-total correlation)	No evidence of separate dimensions when it is assumed that there are or vice versa
Evidence for relations to other variables		
Exploring how scores relate to other measures of similar and dissimilar constructs	Demonstration of moderate to high correlations with scores from other measures of similar constructs	Comparison to measures with unknown or low reliability because a low correlation could reflect measurement error and not true ability
	Demonstration of higher scores with increasing expertise	
	Demonstration of lower procedural complication rates in examinees who are successful on the examination	
Evidence relating to consequences		
Considering the intended and unintended consequences of the examination.	Documentation of rationale for standard-setting procedures	Arbitrary pass/fail decisions
	Provision of feedback and remediation opportunities to trainees	

procedural competence will be presented (Appendix 1, http://links.lww.com/SIH/A242).

SOURCES OF VALIDITY EVIDENCE

Content Evidence

When designing an assessment instrument, one must consider what one wishes to assess. For most assessments, including those that purport to measure the ability to perform procedures, it is not possible to assess the entire content domain, and so, it is important to decide which, and how many, skills should be evaluated. To support content validity, one must present evidence that the construct being assessed is adequately represented on an assessment.[17] Factors that contribute to content validity evidence include the following: the quality of the test items, writer qualifications, the test specifications (ie, the blueprint), and the scoring rubric.[13]

To support the validity argument, the use of content experts to develop test items (ie, the cases or questions used on a test) is essential. It is important to ensure that the selected content experts are credible and have the necessary qualifications (ie, clinicians with recognized expertise in performing the procedures of interest). It is preferable to include multiple content experts who can offer various perspectives and who represent different stakeholder groups. For example, if assessing intubation skills, one might involve clinicians from anesthesia, emergency medicine, and internal medicine. Test items should be reviewed to identify potential flaws that might confuse or cue examinees.

Blueprinting refers to the process by which the content areas to be covered on an examination are defined. Content experts can support the content validity argument by ensuring that the assessment blueprint reflects educational goals and objectives.[18] The need to ensure that a representative amount of content is included must be balanced with feasibility and practical concerns. In assessments of knowledge, it is often not possible to assess examinees on all of the content covered in a curriculum, and so, a representative

© 2015 Society for Simulation in Healthcare

Copyright © 2015 by the Society for Simulation in Healthcare. Unauthorized reproduction of this article is prohibited.

sample of questions is chosen. However, if the sample is too small (at the extreme, imagine a multiple-choice examination with only 1 question), there is a risk of construct underrepresentation, meaning that the domain of interest is not adequately assessed. This is perhaps even more relevant for an assessment of technical skills, because the ability to perform 1 procedure, such as abscess drainage, is not necessarily predictive of one's ability to perform another procedure, such as suturing of a surgical incision. Therefore, to avoid the threat of construct underrepresentation, one must ensure that there are a sufficient number of content-relevant observations on which the assessment is based.[19]

Another consideration when creating a blueprint for an assessment of procedural ability is whether or not to assess skills in context.[20] Successful completion of a medical procedure often requires more than just technical ability (eg, obtaining consent, working as part of team), and incorporating other skill domains can increase realism. As such, some assessments may include hybrid models (ie, a combination of a simulated patient and a partial task model), and others may include allied health professionals or other team members.[20–23] This allows for the creation of cases that more closely mimic reality and may allow for the assessment of the nontechnical skills required when performing procedures, such as communication, collaboration, or professionalism. For example, examinees may be assessed on their ability to suture a laceration while interacting with an anxious simulated patient, or they may need to intubate an unstable patient while managing conflict within a team during a simulation. However, if more than 1 skill is being assessed (eg, technical and communication skills), one must consider how scores will be combined and whether or not measures are compensatory (ie, whether or not strong performance in 1 area can compensate for weaknesses in other areas).

The development of an appropriate scoring instrument is an essential component of the validity argument.[24] Including content experts with knowledge of assessment principles can help ensure that instruments are well constructed. In the case of procedural competence, expert judgment is required to determine which components of a given procedure should be assessed, as well as how they should be assessed (ie, using checklists and/or rating scales).

Procedure-specific checklists are often used when simple yes/no judgments are required regarding the individual steps in a procedure (eg, identified landmarks, obtained synovial fluid, etc). This level of granularity can be useful in guiding raters (because all they need to assess is whether or not an item on the checklist was done) and can allow for the provision of specific feedback on areas of weakness to trainees. Although checklists are often used because they offer an "objective" assessment, they have also been criticized for rewarding a rote approach in which trainees may receive high scores despite missing key steps or even committing egregious errors (such as sterility breaches), which presents a threat to validity.[25] Furthermore, although checklists can reward thoroughness, which may be desirable if all steps in a procedure are essential, they may not capture errors related to the timing or order of steps (eg, creation of a sterile field should be performed before, not after, an incision is made).

Global rating scales (GRSs) are often used either alone or in conjunction with checklists and offer raters an opportunity to assess skills along a continuum (eg, demonstrated good procedural flow), make judgements about complex tasks in which the timing or sequence of actions is important, and assess overall competence. Global rating scales have also been shown to be sensitive to performance differences between experts and novices.[26]

Despite differences of opinion as to whether checklists or GRSs are better, there is evidence that well-structured checklists and well-structured GRSs often produce similar results.[27,28] Regardless of whether checklists or GRSs are used, the scoring criteria must be clear. Careful wording of anchors can help raters distinguish between different levels of competence. In the Objective Structured Assessment of Technical Skills, a performance-based examination testing surgical skills, 7 components of technical skill are assessed using a different 5-point GRS for each.[29] For example, in the rating of "flow of operation," anchors include "frequently stopped operating and seemed unsure of next move," "demonstrated some forward planning with reasonable progression of procedure," and "obviously planned course of operation with effortless flow from one move to the next." Finally, a rationale should be provided for how scores from checklists will be aggregated, how GRS items (if there is more than one) will be combined, and how summary scores will be tallied.

Response Process Evidence

Once the content of the assessment has been determined, one must consider how best to collect data about a trainee's skills. Response process validity evidence demands that assessment data is collected with minimal error.[13] Response process evidence can be broken down into factors related to responses from examinees and raters and those related to data integrity.

For the assessment of procedural competence, where performance-based examinations are often used, one must ensure that examinees are familiar with the examination format (eg, the use of partial task models alone vs. hybrid models, time constraints, etc). They should be provided an opportunity to familiarize themselves with any equipment used, particularly if it differs from the equipment normally available to them. It is also important for them to understand the limitations of the equipment used. For example, when using a partial task model for central venous catheter insertion, examinees should be alerted to the lack of visual cues that would normally be present on a real patient (such as blood color and pulsatility). Examinees should receive an orientation, and instructions and prompts during the examination must be clear. For example, examinees need to know whether they should simply assume the procedure is indicated and proceed or whether they should take steps to ensure that they have obtained consent and ruled out contraindications to the procedure.

When employing raters, it is important to ensure that they understand the instruments used and that they are interpreting the scales as intended. Because human raters demonstrate inherent variability due to preconceived notions and rating characteristics (eg, stringency) that influence their

Copyright © 2015 by the Society for Simulation in Healthcare. Unauthorized reproduction of this article is prohibited.

judgments,[30,31] steps should be taken to reduce these potential sources of error. Rater training has been suggested as 1 method to reduce the potential impact of scorer bias (such as the halo effect). However, the success of rater training has been mixed.[32] For circumstances where rater variability is an issue, 1 helpful strategy involves using multiple different raters for multiple tasks and averaging performance more than the performance domain.[33,34]

If standardized patients or simulators are used in the assessment of procedural ability, it is important to ensure conformity (ie, by training the actors and using comparable equipment between examinees) so that all trainees have a comparable experience.[35] Pilot testing can also be used as part of training to ensure uniformity.

Similarly, if skills have been taught using a specific model or simulator, one must consider the impact of assessing trainees on a different model. Although most studies would suggest that trainees seem to be able to transfer their skills regardless of differences in the type of simulation used,[36–39] there can be an issue of test fairness, an important validity concern, when examinees are provided with different testing experiences.

Quality assurance measures, such as ensuring accurate data collection and entry, also help provide evidence of response process evidence. Using experienced staff to monitor data entry during a performance-based examination and performing audits of data entry are useful practices.[40]

A threat to response process validity relates to cheating. Issues related to examination security and cheating can adversely affect one's interpretation of scores because examinees with previous access to assessment materials (eg, a list of which procedures will be evaluated or a copy of the rating instrument) may be unduly advantaged.[41] Ideally, one must take steps to ensure that all examinees have access to the same information about the assessment.

Internal Structure Evidence

After the administration of an assessment, an analysis of scores can provide further validity evidence. Evidence for internal structure relates to the psychometric characteristics of the instrument, such as reliability and generalizability.[24] Reliability refers to the reproducibility of assessment results and may be estimated using measures of internal consistency such as Cronbach alpha, interrater agreement, or test-retest agreement.[42,43] High reliability suggests that the results of the test are likely to be reproducible if the test was repeated, meaning that examinees who score high on their ability to perform procedures today are likely to perform similarly well if tested tomorrow. In general, reliability can be increased by increasing testing time or the number of test items (ie, questions or cases).[44] Reliability is a necessary component of the validity argument; if an assessment includes measures that are not reliable, then scores resulting from that assessment will have little meaning.

Although the goal of an assessment is to estimate true ability as closely as possible, there is always some degree of measurement error. Various sources of error can be estimated using a generalizability (G) analysis, which considers the simultaneous impact of multiple sources of error variance and could include potential interactions between raters, standardized patients, examinees, and items.[45–47] Once these sources of error are identified, steps can be taken to limit their influence (eg, the number of assessment tasks could be increased or rater training could be enhanced). Further analyses could include studies to identify potential systematic biases (eg, using a differential functioning analysis). From a reproducibility perspective, one needs to ensure that differences in scores are based mainly on differences in trainees' ability to perform procedures, rather than from some other source such as the choice of rater.

Another source of internal structure evidence is related to dimensionality. Often, when designing scoring instruments, the assumption is that some of the items will measure 1 skill or dimension, and another set of items will measure a different skill or dimension. Before reporting results that break the scores down into these dimensions, it is important to determine whether there is any evidence that the multiple dimensions really exist or whether the entire instrument really just measures 1 skill. For example, if 10 items from a rating scale are found to all be highly correlated (ie, >0.9) then one might ask whether it is necessary to include all 10 items, or if fewer items can be used to derive the same information. A factor analysis (ie, a statistical analysis used to uncover groups of related items on a test), or even a simple analysis of correlations across items, can provide evidence related to the underlying dimensions being tested.

The quality of items used may also be inferred by calculating item-total correlations (ITCs). Item-total correlation is a measure of the strength of the association between the total examination score and an item (ie, question or case). A low ITC (ie, usually defined as <0.2) suggests that the item is measuring something other than the intended construct (eg, communication skills rather than procedural ability). As such, a careful review of items with low ITCs is warranted.

Evidence for Relations to Other Variables

Relations between the scores from a particular assessment instrument and other variables can provide further validity evidence.[48] One would expect that scores from 2 instruments attempting to measure the same construct should be highly correlated. Similarly, scores should not correlate strongly (or correlate negatively) with scores from assessment instruments designed to measure different constructs. For example, one might not expect scores from an assessment of communication skills to correlate strongly with scores from an assessment of technical ability. When developing an assessment instrument to measure procedural competence, it is important to ensure that it successfully measures that construct. Perhaps just as important, one must question whether or not it adds information to what can be assessed using more traditional methods (eg, the use of logbooks or in-training evaluations), especially because many simulation-based tools designed to assess procedural competence are costly to implement.[49]

As an example, one might compare scores obtained from a bench examination assessing procedural competence to evaluation forms of trainees completed by preceptors during a surgery rotation. One would expect the results to

Copyright © 2015 by the Society for Simulation in Healthcare. Unauthorized reproduction of this article is prohibited.

be positively correlated, and this relationship should be stronger than scores from unrelated measures (eg, a written test of knowledge). However, if scores were found to be very highly correlated, one might argue that an examination of procedural competence is superfluous and not adding much information about the trainee. If the correlation was negative, one might question the validity of one or both of the sets of scores, because they are purportedly both measuring similar constructs. One could also compare scores to surrogate measures of competence, such as speed or economy of motion, expecting those with higher scores to be more efficient.[50]

When one obtains an unexpectedly low correlation between measures of like constructs, one must consider whether or not this is because the measures are truly unrelated, or if one of both of the tests scores is not sufficiently reliable. That is, if scores on an assessment have poor reliability, it will place a limit on the magnitude of the association with the scores on another assessment. To avoid misinterpreting the strength of the associations, disattenuated correlations (ie, correlations that are adjusted for the known reliabilities of examinations) should be used when possible.[51] When the reliability of scores obtained from 1 or more instruments is not known, correlations should be interpreted with caution.

Comparing the performance of practitioners with different amounts of training or experience is also a potential source of validity evidence.[52] If a tool is measuring the intended construct, one would expect more experienced trainees/practitioners to perform better than novices. For example, staff surgeons and senior surgical residents would be expected to have higher scores on an assessment of technical ability than junior residents and medical students. If this cannot be demonstrated, then the measurement tool may be flawed (eg, rewarding thoroughness more than competence).

The issue of transferability may also be important when evaluating evidence related to relations to other variables. If trainees are successful in meeting minimal criteria when being assessed in a simulated environment, one must consider whether this skill translates to real patients. When simulators, mannequins, or bench models are used, one must question whether or not examinee performance of the skill is likely to be transferable to human subjects. Learning transfer can be defined as "the application of skills and knowledge learned in 1 context to another context."[53] An examinee's ability to perform a procedure on a partial task model may or may not translate into competence to perform that procedure on a patient, which calls into question the predictive validity of the results from that assessment method.

Intuitively, performance assessed using high-fidelity simulations, such as those using virtual reality, would seem to be more likely to translate to real-life situations. However, even low-fidelity simulations have been shown to be useful, at least in training, and many studies would indicate that there is relatively little advantage in using high- rather than low-fidelity simulations.[37,38,54,55] A critical review of simulation-based learning highlighted several examples of positive translational outcomes, such as a reduction of procedure-related complications in patients, providing further evidence that learning transfer occurs in SBME.[10] Although there are relatively few studies that address real outcomes, a recently published systematic review of the literature did identify a positive correlation between SBME assessment scores and patient outcomes in the workplace.[56] When possible, to support the validity argument, researchers should attempt to gather evidence that the skills shown in the simulated environment translate to the "real" world.

Evidence Relating to Consequences

Validity evidence related to consequences refers to the impact of the assessment itself. This may include both intentional and unintentional consequences as they relate to trainees, educators, and perhaps even patients.[57]

For the learner, the intended consequences might include providing an incentive for learning (ie, the test motivates them to study material), receiving feedback to promote continuous learning (such as in formative assessment), and providing them with an endorsement of their qualifications (such as in summative assessment).

Educators may also be interested in the ability of the assessment to discriminate between trainees who are competent, or fit to practice, and those who are not. For high-stakes examinations, such as with board certification, it is especially important to avoid false-positives (ie, passing examinees who do not have adequate knowledge or skills) and false-negatives (ie, failing examinees who are competent). An important source of validity evidence, therefore, rests with the accuracy and process of making pass/fail decisions. This typically includes a description for defensible standard-setting procedures.[58] The rationale for standard-setting procedures (ie, how pass/fail decisions are made) should be explicit. The choice of standard-setting method will depend on several factors, including the purpose of the examination (eg, formative vs. summative), the number of examinees, the type of assessors used (eg, physician examiners vs. trained assessors), and available resources.[59,60] One must also consider whether or not an examinee can pass a given assessment with a high mark if they have missed 1 or more critical steps in a procedure or if they have committed an egregious error (eg, withdrawing cerebrospinal fluid during a lumbar puncture).

An important consequence to consider is how assessment decisions can influence patient care. It seems intuitive that trainees who have been assessed and deemed competent to perform a procedure, even if primarily based on simulation models, would have better patient outcomes than those who have not been assessed or who have been assessed and deemed incompetent. However, when considering consequential validity evidence, it is important to consider that if the categorization of trainees as being competent is based on flawed assessments (ie, those who are lacking in sufficient validity evidence), this could lead to unqualified physicians performing procedures on real patients.

Another example of unintended consequences is that the assessment may uncover weaknesses in a curriculum. Once identified, educators may need to grapple with ways

Copyright © 2015 by the Society for Simulation in Healthcare. Unauthorized reproduction of this article is prohibited.

to address these weaknesses, which could prove challenging in an era in which trainees have restricted work hours and presumably fewer opportunities to perform procedures. This can prove problematic and limit the acceptability of the assessment (eg, if there is a high failure rate because examinees did not receive adequate instruction). There are also implications for educators who must deal with trainees who are not achieving procedural competence. In all likelihood, significant resources will need to be invested in remediating unsuccessful examinees. For trainees, the consequences of failing an examination may be dependent on remediation processes (ie, if they fail an examination but have no or little opportunity to improve due to limited access to procedural training).

The cost of interventions using SBME for the assessment of procedural competence must also be considered as part of consequential validity. Simulation-based medical education is resource intensive and may not be readily available in all training programs. When introducing new SBME curricula, educational leaders may need to justify the cost of its implementation and demonstrate evidence of added value for their programs.

CONCLUSIONS

Interventions using SBME for the assessment of procedural competence can be resource intensive. One way to justify the allocation of resources to SBME assessment is to gather evidence to support the validity of the results and any inferences one wishes to make on the basis of them (eg, competency decisions). There is clearly a need to ensure that procedural competence is being assessed. We should aim to begin assessing procedural skills as rigorously as we assess cognitive skills.

Methods to assess procedural competence range from informal assessments, such as procedure logs or unstructured observational assessments, to highly structured performance-based examinations. Whichever method is used, one should consider the purpose of the assessment and examine the evidence for the validity of the scores obtained from that method. Content, response process, internal structure, relationship to other variables, and the consequences of the assessment can all provide evidence to support the validity of score-based inferences concerning the procedural acumen of those being assessed. Threats to the validity of the assessment need to be considered.

REFERENCES

1. Kohn KT, Corrigan JM, Donaldson MS, eds. Errors in health care: a leading cause of death and injury. In: Kohn KT, Corrigan JM, Donaldson MS, eds. *To Err Is Human: Building a Safer Health Care System.* Washington, DC: National Academy Press; 1999:26–48.
2. Antiel RM, Reed DA, Van Arendonk KJ, et al. Effects of duty hour restrictions on core competencies, education, quality of life, and burnout among general surgery interns. *JAMA Surg* 2013;148:448–455.
3. Sen S, Kranzler HR, Didwania A, et al. Effects of the 2011 duty hour reforms on interns and their patients: a prospective longitudinal cohort study. *JAMA Intern Med* 2013;173:657–662.
4. Kaboli PJ, Go JT, Hockenberry J, et al. Associations between reduced hospital length of stay and 30-day readmission rate and mortality: 14-year experience in 129 Veterans Affairs hospitals. *Ann Intern Med* 2012;157:837–845.
5. Issenberg SB, McGaghie WC, Petrusa ER, Lee Gordon D, Scalese RJ. Features and uses of high-fidelity medical simulations that lead to effective learning: a BEME systematic review. *Med Teach* 2005;27:10–28.
6. Schaefer JJ 3rd, Vanderbilt AA, Cason CL, et al. Literature review: instructional design and pedagogy science in healthcare simulation. *Simul Healthc* 2011;6:S30–S41.
7. McGaghie WC, Issenberg SB, Petrusa ER, Scalese RJ. A critical review of simulation-based medical education research: 2003-2009. *Med Educ* 2010;44:50–63.
8. Cook DA, Brydges R, Zendejas B, Hamstra SJ, Hatala R. Technology-enhanced simulation to assess health professionals: a systematic review of validity evidence, research methods, and reporting quality. *Acad Med* 2013;88:872–883.
9. Cook DA, Hamstra SJ, Brydges R, et al. Comparative effectiveness of instructional design features in simulation-based education: systematic review and meta-analysis. *Med Teach* 2013;35:e867–e898.
10. McGaghie WC, Issenberg SB, Barsuk JH, Wayne DB. A critical review of simulation-based mastery learning with translational outcomes. *Med Educ* 2014;48:375–385.
11. Cook DA, Zendejas B, Hamstra SJ, Hatala R, Brydges R. What counts as validity evidence? Examples and prevalence in a systematic review of simulation-based assessment. *Adv Health Sci Educ Theory Pract* 2014;19:233–250.
12. Cook D. One drop at a time: research to advance the science of simulation. *Simul Healthc* 2010;5:1–4.
13. Downing SM. Validity: on meaningful interpretation of assessment data. *Med Educ* 2003;37:830–837.
14. American Educational Research Association (AERA), American Psychological Association (APA), National Council on Measurement in Education (NCME). *Standards for Educational and Psychological Testing.* Washington: AERA; 2014.
15. Kane MT. Validating the interpretations and uses of test scores. *J Educ Meas* 2013;50:1–73.
16. Cizek GJ. Defining and distinguishing validity: interpretations of score meaning and justifications of test use. *Psychol Methods* 2012;17:31–43.
17. Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: theory and application. *Am J Med* 2006;119:166.e7–166.e16.
18. Coderre S, Woloschuk W, McLaughlin K. Twelve tips for blueprinting. *Med Teach* 2009;31:322–324.
19. Norman G, Bordage G, Page G, Keane D. How specific is case specificity? *Med Educ* 2006;40:618–623.
20. Ellaway RH, Kneebone R, Lachapelle K, Topps D. Practica continua: Connecting and combining simulation modalities for integrated teaching, learning and assessment. *Med Teach* 2009;31:725–731.
21. Stroud L, Cavalcanti RB. Hybrid simulation for knee arthrocentesis: improving fidelity in procedures training. *J Gen Intern Med* 2013;28:723–727.
22. Kneebone R, Nestel D, Yadollahi F, et al. Assessing procedural skills in context: exploring the feasibility of an Integrated Procedural Performance Instrument (IPPI). *Med Educ* 2006;40:1105–1114.
23. Pugh D, Hamstra SJ, Wood TJ, et al. A procedural skills OSCE: assessing technical and non-technical skills of internal medicine residents. *Adv Health Sci Educ Theory Pract* 2015;20:85–100.
24. Downing SM, Haladyna T. Validity and its threats. In: Downing SM, Yudkowsky R, eds. *Assessment in Health Professions Education.* 1st ed. New York: Routledge; 2009:21–55.
25. Cunnington JP, Neville AJ, Norman GR. The risks of thoroughness: reliability and validity of global ratings and checklists in an OSCE. *Adv Health Sci Educ Theory Pract* 1996;1:227–233.
26. Hodges B, Regehr G, McNaughton N, Tiberius R, Hanson M. OSCE checklists do not capture increasing levels of expertise. *Acad Med* 1999;74:1129–1134.
27. Boulet JR, van Zanten M, de Champlain A, Hawkins RE, Peitzman SJ. Checklist content on a standardized patient assessment: an ex post facto review. *Adv Health Sci Educ Theory Pract* 2008;13:59–69.

Copyright © 2015 by the Society for Simulation in Healthcare. Unauthorized reproduction of this article is prohibited.

28. Ilgen JS, Ma IW, Hatala R, Cook DA. A systematic review of validity evidence for checklists versus global rating scales in simulation-based assessment. *Med Educ* 2015;49:161–173.

29. Reznick R, Regehr G, MacRae H, Martin J, McCulloch W. Testing technical skill via an innovative "bench station" examination. *Am J Surg* 1997;173:226–230.

30. Wood TJ. Exploring the role of first impressions in rater-based assessments. *Adv Health Sci Educ Theory Pract* 2014;19:409–427.

31. Williams RG, Klamen DA, McGaghie WC. Cognitive, social and environmental sources of bias in clinical performance ratings. *Teach Learn Med* 2003;15:270–292.

32. Cook DA, Dupras DM, Beckman TJ, Thomas KG, Pankratz VS. Effect of rater training on reliability and accuracy of mini-CEX scores: a randomized, controlled trial. *J Gen Intern Med* 2009;24:74–79.

33. Eva KW, Rosenfeld J, Reiter HI, Norman GR. An admissions OSCE: the multiple mini-interview. *Med Educ* 2004;38:314–326.

34. Boulet JR, McKinley DW. Criteria for a good assessment. In: McGaghie WC, ed. *International Best Practices for Evaluation in the Health Professions.* Oxford, UK: Radcliffe Publishing Ltd.; 2013:19–43.

35. Pugh D, Smee S. Guidelines for the Development of Objective Structured Clinical Examination (OSCE) Cases. Medical Council of Canada, 2013. Available at: http://mcc.ca/wp-content/uploads/osce-booklet-2014.pdf. Accessed April 9, 2015.

36. Hatala R, Issenberg SB, Kassen B, Cole G, Bacchus CM, Scalese RJ. Assessing cardiac physical examination skills using simulation technology and real patients: a comparison study. *Med Educ* 2008;42:628–636.

37. Anastakis DJ, Regehr G, Reznick RK, et al. Assessment of technical skills transfer from the bench training model to the human model. *Am J Surg* 1999;177:167–170.

38. Matsumoto ED, Hamstra SJ, Radomski SB, Cusimano MD. The effect of bench model fidelity on endourological skills: a randomized controlled study. *J Urol* 2002;167:1243–1247.

39. de Giovanni D, Roberts T, Norman G. Relative effectiveness of high- versus low-fidelity simulation in learning heart sounds. *Med Educ* 2009;43:661–668.

40. Boulet JR, McKinley DW, Whelan GP, Hambleton RK. Quality assurance methods for performance-based assessments. *Adv Health Sci Educ Theory Pract* 2003;8:27–47.

41. Gotzmann A, de Champlain A, Homayra F, et al. Cheating in OSCEs: The impact of simulated security breaches on OSCE performance. Submitted to Medical Education, August 31, 2015.

42. Downing SM. Reliability: on the reproducibility of assessment data. *Med Educ* 2004;38:1006–1012.

43. Cortina J. What is coefficient alpha? An examination of theory and applications. *J Appl Psychol* 1993;78:98–104.

44. Newble DI, Swanson DB. Psychometric characteristics of the objective structured clinical examination. *Med Educ* 1988;22:325–334.

45. Boulet J. *Generalizability Theory: Basics, Encyclopedia of Statistics in Behavioral Science.* In: Everitt BS, Howell DC, eds. Chichester: John Wiley & Sons Ltd; 2005:704–711.

46. Burns KJ. Beyond classical reliability: using generalizability theory to assess dependability. *Res Nurs Health* 1998;21:83–90.

47. Reznick R, Smee S, Rothman A, et al. An objective structured clinical examination for the licentiate: report of the pilot project of the Medical Council of Canada. *Acad Med* 1992;67:487–494.

48. Campbell DT, Fiske DW. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychol Bull* 1959;56:81–105.

49. Savoldelli GL, Naik VN, Joo HS, et al. Evaluation of patient simulator performance as an adjunct to the oral examination for senior anesthesia residents. *Anesthesiology* 2006;104:475–481.

50. Grober ED, Roberts M, Shin EJ, Mahdi M, Bacal V. Intraoperative assessment of technical skills on live patients using economy of hand motion: establishing learning curves of surgical competence. *Am J Surg* 2010;199:81–85.

51. Downing S. *Statistics of testing, Assessment in Health Professions Education.* 1st ed. In: Downing SM, Yudkowsky R, eds. New York: Routledge; 2009:93–117.

52. LeBlanc VR, Tabak D, Kneebone R, Nestel D, MacRae H, Moulton CA. Psychometric properties of an integrated assessment of technical and communication skills. *Am J Surg* 2009;197:96–101.

53. Hamstra SJ, Dubrowski A, Backstein D. Teaching technical skills to surgical residents: a survey of empirical research. *Clin Orthop Relat Res* 2006;449:108–115.

54. Hamstra SJ, Brydges R, Hatala R, Zendejas B, Cook DA. Reconsidering fidelity in simulation-based training. *Acad Med* 2014;89:387–392.

55. Norman G, Dore K, Grierson L. The minimal relationship between simulation fidelity and transfer of learning. *Med Educ* 2012;46:636–647.

56. Brydges R, Hatala R, Zendejas B, Erwin PJ, Cook DA. Linking simulation-based educational assessments and patient-related outcomes: a systematic review and meta-analysis. *Acad Med* 2015;90:246–256.

57. Haertal E. Gettting the help we need. *J Educ Meas* 2013;50:84–90.

58. Downing SM, Tekian A, Yudkowsky R. Procedures for establishing defensible absolute passing scores on performance examinations in health professions education. *Teach Learn Med* 2006;18:50–57.

59. Boulet JR, Murray D, Kras J, Woodhouse J. Setting performance standards for mannequin-based acute-care scenarios: an examinee-centered approach. *Simul Healthc* 2008;3:72–81.

60. McKinley DW, Norcini JJ. How to set standards on performance-based examinations: AMEE Guide No. 85. *Med Teach* 2014;36:97–110.

Copyright © 2015 by the Society for Simulation in Healthcare. Unauthorized reproduction of this article is prohibited.