

formally adopted by the American Educational Research Association, American Psychological Association, and National Council on Measurement in Education in 1999,²⁰ and it remains the standard approach by which to evaluate the validity of an instrument's scores.²¹

Because judgments of evidence rest on clear and transparent reporting, it is also important to understand issues related to the reporting of validity evidence. Reviews in medical education have appraised the reporting quality of studies evaluating educational interventions,^{22–24} but we are not aware of appraisals of reporting quality for studies evaluating assessment tools.

To address these gaps, we sought to identify and summarize the validity evidence and reporting quality for all studies of technology-enhanced simulation-based assessment involving health professions learners. We did this by conducting a systematic review in which we evaluated the prevalence of validity evidence,²⁰ potential methodological biases and limitations,^{25,26} and reporting quality.^{27,28} We sought to answer the following questions:

1. What are the characteristics of tools to assess learning outcomes (knowledge, skills, attitudes) using technology-enhanced simulation?
2. What validity evidence has been reported for these assessments?
3. What is the methodological quality and reporting quality of the studies from which this evidence derives?

Method

This review was planned, conducted, and reported in adherence to PRISMA standards of quality for reporting systematic reviews.²⁹ We conducted this review of simulation-based assessment concurrently with a review of simulation-based training. Although these reviews addressed different questions and employed distinct inclusion criteria, some details of study identification and selection have been published previously.⁸

Evaluating the validity of education assessments

Before conducting our search of the literature, we determined the criteria and instruments we would use to evaluate the articles. We coded the prevalence of

each of the five evidence sources noted above: content, response process, internal structure, relations with other variables, and consequences (see Table 1 for definitions). For internal structure and relations with other variables, we counted separately several distinct elements (see Table 1). Kane¹⁸ has extended Messick's¹⁹ framework by emphasizing the importance of a systematic approach to validation, including a carefully articulated validity argument, so we coded for the presence of validity arguments in the planning and interpretation of each study. We focused on assessment and did not include evidence regarding the "validity" of training activities.

Evaluating method and reporting quality for assessment studies

We found no instruments for the appraisal of studies evaluating educational assessments. However, we identified three evidence-based instruments for appraising the methodological or reporting quality of studies of clinical diagnostic tests.^{25,27,28} The paradigm of clinical diagnosis applies readily to educational assessment, inasmuch as the intent of assessment is to make judgments (i.e., diagnoses) about learners for the purpose of making

Table 1

Definitions and Prevalence of Validity Evidence in a 2011 Systematic Review of Technology-Enhanced Simulation for Assessment

Evidence element	Definition*	Prevalence, no. (%)	
		All studies, N = 417	Studies reviewed in detail, N = 217
Content†	Description of steps taken to ensure that test content reflects the construct it is intended to measure	142 (34)	137 (63)
Response process	Analysis of raters' thoughts/actions while scoring behavior, test security, quality control	14 (3)	13 (6)
Internal structure—reliability	Reproducibility of scores across ...		
Interrater reliability	... different raters	124 (30)	117 (54)
Interstation reliability	... different stations or tasks	40 (10)	40 (18)
Test–retest reliability	... different versions of the test	22 (5)	19 (9)
Internal consistency	... different items on the test	46 (11)	46 (21)
Any	... any facet of variation	163 (39)	153 (71)
Internal structure—item analysis	Empiric evaluation of scoring, interitem correlation, item discrimination	40 (10)	40 (18)
Internal structure—factor analysis	Exploratory or confirmatory factor analysis	2 (0.5)	2 (1)
Relations with other variables—separate measure	Association with separate measure with a hypothesized relationship (positive or negative) with test scores	128 (31)	102 (47)
Relations with other variables—learner characteristic	Association with training level (expert/novice) or status (trained/untrained)	306 (73)	168 (77)
Consequences	Impact, beneficial or harmful, of the assessment itself	20 (5)	20 (9)

*Cook and Beckman²¹ provide more detailed definitions.

†The prevalence of content evidence here is lower than that coded using the Medical Education Research Study Quality Instrument (see Table 5) because new evidence was required in this phase of coding (as compared with citing evidence previously published).

decisions regarding, for example, mastery or needed improvements. One important difference in research design is that studies evaluating clinical tests typically employ an independent gold standard, whereas gold standards are rarely available for educational assessments. However, we found substantial overlap in most reporting and methodological domains.

The Standards for Reporting Diagnostic Accuracy (STARD)²⁷ were published in 2003, for the purpose of “improv[ing] the quality of reporting of studies of diagnostic accuracy.” Whereas some items on this checklist refer to studies making comparison with a reference test (gold standard), most items apply generally to any study of a diagnostic test. The Guidelines for Reporting Reliability and Agreement Studies (GRRAS),²⁸ published in 2011, complement the STARD by emphasizing items specific to reliability studies.

To evaluate study biases, we used the revised Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2).²⁵ The purpose of this seven-question tool is to “assess the quality of primary diagnostic accuracy studies” in the context of a systematic review. Three domains (participant selection, index test, and reference test) each have questions on possible bias (systematic flaws that distort study results) and applicability to the review question. A fourth domain evaluates bias in study flow. We did not code the participant selection-applicability question (i.e., the match between study participants and the review question) because all studies, per inclusion criteria, enrolled health professions learners.

We also evaluated study methods using the Medical Education Research Study Quality Instrument²⁶ (MERSQI), which was developed to appraise the methodological quality of any quantitative research study.

Study eligibility

We included original research studies published in any language that had as a stated purpose the evaluation of technology-enhanced simulation for assessing health professions learners at any stage in training or practice. We made no restrictions based on study design or validity evidence reported. We defined technology-enhanced simulation as an educational tool or device with which the

learner physically interacts to mimic an aspect of clinical care for the purpose of teaching or assessment.⁸

Study identification

Our search strategy has been previously published in full.⁸ To summarize briefly, we searched MEDLINE, EMBASE, CINAHL, PsycINFO, ERIC, Web of Science, and Scopus for relevant articles using a search strategy developed by an experienced research librarian. The search included terms related to the topic (simulat*, mannequin, virtual, etc.), population (education medical, education nursing, students health occupations, etc.), and assessment (assess*, evaluat*, valid*, reliab*, etc.). We used no beginning date cutoff, and the last date of search was May 11, 2011. We supplemented this search by examining the entire reference list from several published reviews^{9–11,14–16} and all articles published in two journals devoted to health professions simulation (*Simulation in Healthcare* and *Clinical Simulation in Nursing*).

Study selection

We worked independently and in duplicate to screen all candidate studies for inclusion, beginning with titles and abstracts and proceeding to the full text of studies judged eligible or uncertain. We resolved conflicts by consensus. Chance-adjusted interrater agreement for study inclusion, determined using intraclass correlation coefficient (ICC), was 0.72.

Data extraction and synthesis

We developed a data abstraction form through iterative testing and revision. We abstracted data independently and in duplicate for all variables (see Tables 1 and 2 for definitions), resolving conflicts by consensus. We abstracted data in two levels: basic and detailed. For all studies, we abstracted information on the number and training level of learners, clinical topic, study design, outcomes (reaction, knowledge, skills [distinguished as time, process, and product skills[†]], behaviors, and patient effects, as previously detailed),⁸ validity evidence (as above), and methodological quality (using the MERSQI). For studies reporting two or more elements of validity evidence, we

abstracted additional details on learners, raters, outcome metrics, validity evidence, study quality (using the QUADAS-2), and reporting quality (using the STARD and GRRAS). We employed the two-level approach for reasons of feasibility, and also to ensure that coded studies had the evaluation of an assessment tool as a central focus (rather than as a small part of a study with a different focus). We would expect that studies reporting more validity evidence are generally better designed and reported overall, and thereby that excluding studies with less evidence most likely overestimates the quality of methods and reporting for the sample as a whole. However, we did not collect empiric data in this regard. ICC for this basic/detailed decision was 0.80.

ICCs for MERSQI variables ranged from 0.51 to 0.84. For validity screening, ICCs ranged from 0.67 to 0.91 except for response process (ICC = 0.34, raw agreement 95%) and consequences (ICC = 0.56), and for QUADAS-2 scores, ICCs ranged from 0.55 to 0.72 except for index test and reference test applicability (ICC = 0.17 and ICC = 0.39, raw agreement 94% and 83%, respectively). For reporting quality, nearly all ICCs were >0.5, and all were >0.3 except prospective/retrospective data collection (ICC = 0.27), rationale for test relationship (ICC = 0.22), and correlation coefficient confidence interval (ICC = 0, raw agreement 99%). ICC values 0.21 to 0.4 are considered “fair,” 0.41 to 0.6 are “moderate,” and 0.61 to 0.8 are “substantial.”³⁰

Most studies employed more than one assessment tool. In such instances, we selected one tool to code on the basis of (in order of priority) (1) the strongest validity evidence, (2) the tool noted in the report title or purpose, (3) a named tool, or (4) the tool measuring the highest outcome.

We summarized the data using counts and, where appropriate, means. We used SAS 9.3 (SAS Institute Inc., Cary, North Carolina) to perform *t* test and chi-square analyses of change over time, with an alpha level of 0.05.

Results

Trial flow and participants

From 10,911 potentially relevant articles, we included 417 studies enrolling 19,075

[†]Time skills refers to the time required to perform a task, process skills refers to performance during a task (e.g., global ratings or minor errors), and product skills refers to the final result (e.g., successful completion, major complication, or the quality of the final product).⁸

Table 2

Criteria and Prevalence of Reporting Quality as Determined by the Standards for Reporting Diagnostic Accuracy (STARD) and the Guidelines for Reporting Reliability and Agreement Studies (GRRAS) in a 2011 Systematic Review of Technology-Enhanced Simulation for Assessment

Study component	STARD Item*	GRRAS Item*	Reporting element (operational definition)	No. possible [†]	Prevalence, no. [†] (%)
Title/abstract					
	1. Identified as study of diagnostic accuracy	1. Identified that reliability was investigated	Title or abstract identifying the study as an evaluation of the validity, reliability, or diagnostic accuracy of an assessment tool	217	154 (71)
			Title or abstract identifying the study as focused on assessment, but not as a study of validity, reliability, or diagnostic accuracy	217	47 (22)
Introduction					
	2. Research question		Explicit question, purpose, or hypothesis	217	186 (86)
			Proposed validity argument (strategy for interpreting validity evidence to be presented)	217	139 (64)
		2. Instrument name, description	Description of index test task	217	172 (79)
		5. Existing evidence for this instrument	Critical review of evidence relevant to assessment of that construct	217	135 (62)
Method					
	3. Study population: eligibility, setting	3. Study population	Trainee population (eligibility criteria)	217	58 (27)
			Setting (educational [e.g., simulation laboratory] versus clinical)	217	108 (50)
	4. Participant recruitment		Identification of eligible trainees (any method defined)	217	127 (59)
	5. Participant sampling: consecutive, random	7. Sampling method	Sampling strategy (any method defined)	217	38 (18)
	6. Data collection: prospective or retrospective		Prospective or retrospective data collection	217	20 (9)
	7. Reference standard: definition, rationale		Rationale for relationship between index and reference test	102	52 (51)
	8. Index test and reference standard: methods	8. Measurement procedures described	Methods/procedures for index test	217	192 (88)
			Methods/procedures for reference test.	102	90 (88)
	9. Index test and reference standard: classification		Passing standard	217	33 (15)
	10. Raters: number, training	4. Rater population	Rater population (eligibility criteria)	159	12 (8)
			Rater training (done or not done)	159	69 (43)
		11. Rater number	Rater total number	159	139 (87)
		12. Rater characteristics	Rater specialty	159	101 (64)
	11. Raters: blinded	9. Raters independent	Raters blinded to trainee (done or not done)	159	86 (54)
			Raters blinded to other raters (done or not done)	135	86 (64)
			Raters blinded to results of reference test (done or not done)	95	30 (32)
	12. Statistical methods—accuracy: defined	10. Statistical methods	All statistical methods defined: comparisons among groups or correlation	198	190 (96)
	13. Statistical methods—reliability: defined		All statistical methods defined: reliability	153	135 (88)
		6. Sample size calculations (planned)	Sample size calculations	217	19 (9)

(Table continues)

Table 2
(Continued)

Study component	STARD Item*	GRRAS Item*	Reporting element	No. [†]	Prevalence, no. [†] (%)
Results					
	14. Study dates		Study dates	217	53 (24)
	15. Participant demographics	11. Participant number	Trainee number enrolled	217	205 (94)
		12. Characteristics, participants	Trainee training level	217	165 (76)
	16. Participants eligible, not enrolled; flow diagram		Trainee number eligible	217	30 (14)
			Flow diagram	217	4 (2)
	17. Time and events between index and reference tests		Time interval between index and reference test	102	74 (73)
	18. Severity of disease in population		Trainee baseline proficiency: objective measurement	217	4 (2)
			Trainee baseline proficiency: prior experience with that task	217	63 (29)
	19. Distribution of test results		Central tendency (mean, median) and variability (standard deviation, range) for scores	217	171 (79)
			Central tendency (mean, median) without variability	217	22 (10)
			Figure (scatter plot) or table (contingency table)	102	31 (30)
	20. Adverse events		Consequences of testing, adverse or beneficial	217	6 (3)
	21. Estimates of accuracy and statistical uncertainty		Estimate of accuracy (correlation coefficient or other)	217	92 (91)
			Receiver operating characteristic (ROC) curve, sensitivity, or specificity of test	217	2 (0.5)
			Confidence intervals for accuracy estimates	92	2 (2)
	22. Indeterminate and outlier results: how handled		Scoring process described	217	159 (73)
			Indeterminate and outlier results considered in scoring	217	21 (10)
	23. Variability across subgroups of participants or raters		Subgroup analyses interpreted as relating to score validity	217	1 (0.5)
	24. Reproducibility	13. Reliability, including uncertainty	Reliability (any)	217	153 (71)
			Confidence intervals for reliability estimates	153	16 (10)
Discussion	25. Clinical applicability	14. Practical relevance	Concluding validity argument (interpretation of validity evidence)	217	168 (77)

* Numbers refer to item number in original STARD²⁷ or GRRAS²⁸ statement.

[†] N = 217 studies for most items. There were 159 studies with human raters, 135 studies with two or more raters, 102 studies with a reference test, and 95 studies with both human raters and a reference test. One hundred fifty-three studies reported reliability, and 198 reported either a correlation between two measures or a contrast between two or more groups.

trainees (median 30 trainees per study, interquartile range 20–50); see Figure 1 for details. Four of these were published in a language other than English. Forty-one percent of the articles (N = 172) were published in or after 2008. Table 3 summarizes study characteristics. Because we found so many studies, we do not reference them all in this report. However, Supplemental Digital Appendix 1 provides a complete list of studies,

Supplemental Digital Table 1 provides key coding results, and Supplemental Digital Table 2 lists instruments by clinical topic (<http://links.lww.com/ACADMED/A130>). We coded all 417 included studies for tool characteristics, validity evidence, and MERSQI criteria. We coded 217 studies (those reporting two or more elements of validity evidence) in greater detail using the QUADAS-2, STARD, and GRRAS criteria.

Three hundred fifty studies (84%) involved physicians at some stage in training, including 281 (67%) that involved postgraduate physician trainees (residents), 208 (50%) that involved practicing physicians, and 115 (28%) that involved medical students (some studies included participants from more than one training stage). Twenty-six studies (6%) involved nurses, and 33 studies (8%) involved other

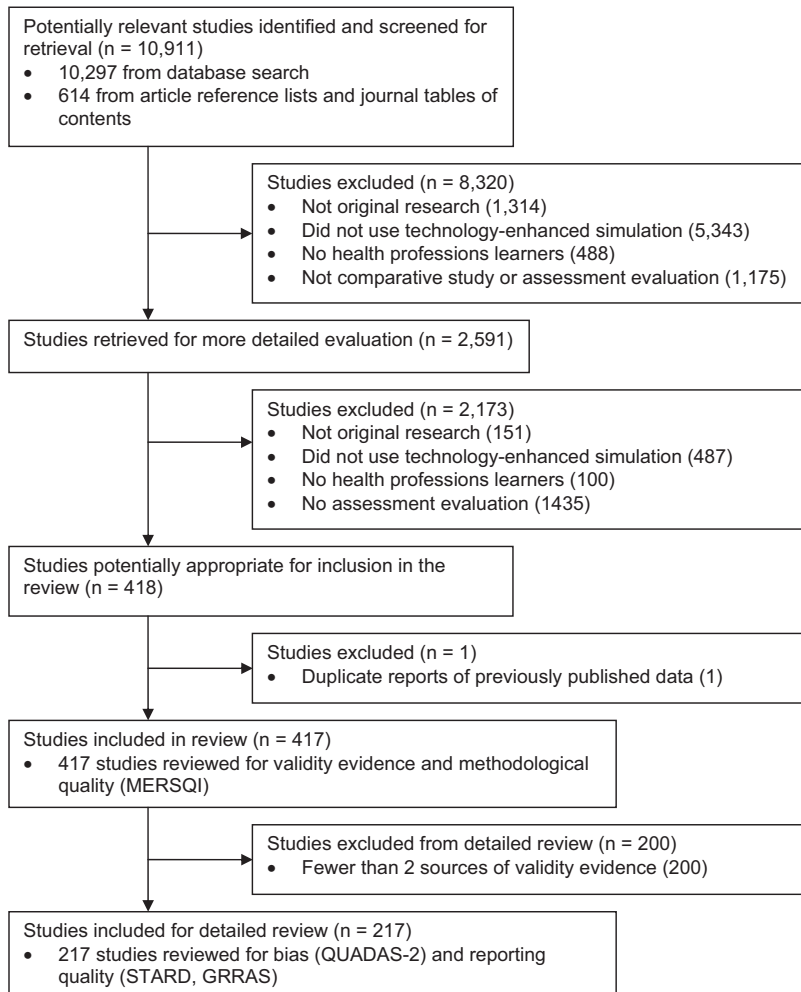


Figure 1 Trial flow for a 2011 systematic review of technology-enhanced simulation for assessment. MERSQI indicates Medical Education Research Study Quality Instrument; QUADAS-2, Quality Assessment of Diagnostic Accuracy Studies version 2; STARD, Standards for Reporting Diagnostic Accuracy; GRRAS, Guidelines for Reporting Reliability and Agreement Studies.

trainees including emergency medical technicians, dentists, and respiratory therapists. We could not quantify precisely how many trainees participated from each category because 115 studies enrolling 4,283 trainees did not clearly define trainee levels (e.g., combining postgraduate and practicing physicians as “experienced”).

Tool characteristics

Studies evaluated the use of technology-enhanced simulations to assess learners in diverse topics, including laparoscopic and open surgery, gastrointestinal and urological endoscopy, anesthesiology, obstetrics–gynecology, and physical examination of the heart, breast, and prostate (see Table 3). By far the most common outcome was process skill, assessed in 356 studies (85%), followed by

time (172 studies, 41%) and product skills (53 studies, 13%). Thirty-one studies (7%) evaluated nontechnical outcomes, such as communication and team leadership.

Table 4 lists the named assessment tools reported five or more times. Aside from the Objective Structured Assessment of Technical Skills (OSATS)³¹ and the McGill Inanimate System for Training and Evaluation of Laparoscopic Skills (MISTELS),³² these common tools involved computerized virtual reality and/or motion tracking. Consistent with this observation, the most commonly used simulator devices were computer-based virtual reality systems, employed in 171 studies (41%), followed by part-task synthetic models (156 studies, 37%) and mannequins (96 studies, 23%). Live animals were used in 14 studies (3%).

Looking at trends over time, the proportionate use of simulators since 2008 is similar to the overall sample, with 72 (42%), 60 (35%), and 38 (22%) of the 172 studies published in or since 2008 involving virtual reality, models, and mannequins, respectively. Seventeen of the 31 studies of nontechnical skills (55%) were published since 2008.

Validity evidence

Table 1 summarizes the validity evidence presented in the 417 articles. By far the most common evidence element was relations with a learner characteristic such as training status (procedural experience or training level), addressed in 306 (73%) studies. One hundred thirty-eight studies (33%) reported no validity evidence other than this. Evidence of content, reliability, and relations with a separately measured variable were each reported in approximately one-third of studies (N = 142, 163, and 128, respectively). The prevalence of content evidence here is lower than that reported for the MERSQI (Table 5) because we required new evidence, whereas the MERSQI credits the presentation of previously published evidence. Response process and consequences evidence were infrequently reported (≤5% each).

Under the category of relations with other variables, 28 studies evaluated associations between simulation-based performance and performance with real patients. These outcomes included measures of procedural time (N = 6), behaviors (instructor ratings of technique, N = 25), and patient effects (rate of procedural success or complications, N = 4). Without exception, these studies showed that higher simulator scores were associated with higher performance in clinical practice.

One hundred eighty-one studies reported only one evidence element, 78 reported two elements, and 139 reported three or more. Nineteen studies reported no substantive evidence, despite having as an aim the evaluation of an assessment tool. Of the 163 studies reporting reliability data, 106 reported one reliability type (most often interrater reliability).

Seventy-five of the 217 studies reviewed in detail (35%) used a “classical” validity framework³³ (content, criterion, and

Table 3

Description of Studies Included in a 2011 Systematic Review of Technology-Enhanced Simulation for Assessment

Study characteristic	Level	No. of studies	No. of participants*
All studies [†]		417	19,075
Participants [‡]	Medical students	115	3,197
	Physicians postgraduate training	281	6,660
	Physicians in practice	208	3,056
	Nurses and nursing students	26	1,318
	Emergency medical technicians and students	4	105
	Dentists and dental students	10	293
	Other	17	163
	Ambiguous / mixed	115	4,283
Clinical topics [§]	Minimally invasive surgery	142	6,144
	Other surgery	81	3,706
	Endoscopy and ureteroscopy	67	2,346
	Resuscitation / trauma training	45	2,922
	Nontechnical skills [¶]	31	2,209
	Anesthesia	31	1,838
	Endovascular procedures	18	659
	Physical examination	12	1,573
	Obstetrics	10	322
	Airway management	9	381
	Dentistry	9	248
	Vascular access	7	1,087
Simulator outcomes [‡]	Knowledge	1	89
	Skill: time	172	6,693
	Skill: process	356	16,403
	Skill: product	53	2,242

*Numbers reflect the number enrolled. Most studies included trainees from more than one level.

[†]See Supplemental Digital Table 1 for details on individual studies, <http://links.lww.com/ACADMED/A130>.

[‡]The number of studies and learners sum to more than the number for all studies because most studies included more than one learner group or reported multiple outcomes, and several fit within more than one clinical topic.

[§]Selected listing of the topics addressed most often. Several other topics were addressed, with lower frequency (data not shown).

[¶]Nontechnical skills include communication, leadership/team management, organization, situational awareness, and decision making.

construct validity) for planning and interpreting data. Another 85 (39%) used a more limited framework, such as “construct” validity alone, and 51 (24%) reported no validity framework. Only 6 (3%) invoked the currently accepted model.²⁰

Looking at trends over time, the number of validity evidence sources has decreased slightly in recent years: Before 2008, each study reported on average 2.14 (SD 1.36) evidence sources; since 2008, the average was 1.98 (SD 1.41). Considering evidence sources separately, the change pre- to post-2008 was less than ± 4 percentage points and not statistically significant ($P > .05$) except for relations with

a separate measure (35% pre-2008, decreasing to 25%; $P = .03$).

Methodological quality

Table 5 summarizes the methodological quality of all 417 articles as evaluated using the MERSQI. Over half ($N = 225$) were single-group, single-assessment (i.e., cross-sectional) studies, whereas another one-third ($N = 152$) employed a one-group pretest–posttest or crossover design. The vast majority ($N = 398$; 95%) employed objective outcome measurements. Thirty-nine studies (9%) made a statistical error in a main analysis.

In our two-level approach, we selected all studies reporting two or more elements of validity evidence for additional coding

of methodological and reporting quality. As seen in Table 5, MERSQI scores for these 217 studies were very similar to the full set, except (as would be expected) for the prevalence of validity evidence.

We evaluated these 217 studies using the QUADAS-2. As shown in Table 5, we found only 25 studies (12%) at low risk of bias in participant selection. High bias was typically due to expert–novice (case–control) comparisons, whereas unclear bias was due to failure to describe the population (data not shown). We also had frequent concerns about the conduct of the index test and reference test ($N = 85$ of 217 [39%] and $N = 34$ of 102 [33%] judged low risk of bias, respectively). Most reference tests aligned reasonably well with the target condition ($N = 83$ [81%] judged low concern about applicability), as did nearly all of the index tests ($N = 205$ [94%] low concern).

Although the STARD criteria reported in Table 2 (discussed below) focus on reporting quality, in abstracting these elements we also coded methodological quality. For example, 86 of 159 studies with human raters reported whether or not raters were blinded to trainee experience (i.e., done or not done)—but blinding was actually done in only 66 (42%) (i.e., it was reported as not done in 20). Raters were blinded to one another in 83 of 135 studies with two or more raters (61%) and blinded to the reference test in 12 of 95 studies with a reference test (13%). Twenty-three studies (14%) employed only one rater per observation. Raters completed special training in 67 of 159 studies (42%). Among the 38 studies reporting a sampling strategy, 20 enrolled the entire available sample (e.g., an entire medical school class), 5 enrolled a random sample, and 13 employed defined inclusion/exclusion criteria.

Reporting quality

Table 2 summarizes reporting quality as measured by the STARD and GRRAS. Nearly all studies ($N = 186$; 86%) reported a focused question, but only 135 (62%) offered a critical review of relevant literature (“cites articles relevant to the topic or study design and critically discusses these articles”),²² and 139 (64%) proposed a plan for interpreting the evidence to be presented (validity argument). Trainee flow was sparsely reported,

Table 4
Commonly Reported Tools for Simulation-Based Assessment in a 2011 Systematic Review of Technology-Enhanced Simulation for Assessment

Tool name	Tool description*	No. of studies, primary assessment method†	No. of participants, primary assessment method†	No. of studies, secondary assessment method‡	Validity evidence§					
					Content	Response process	Internal structure	Relations with other variables: other measure	Relations with other variables: training level	Consequences
OSATS (Objective Structured Assessment of Technical Skills) ³¹	Human rater (TS, surgery): task-specific checklist, multi-item global rating scale, pass/fail rating	27	1,008	7	14	0	24	16	25	1
MISTELS (McGill Inanimate System for Training and Evaluation of Laparoscopic Skills) ³²	Human rater (TS, MIS): 5–7 box trainer tasks; composite score (time with error penalty)	14	806	8	0	2	2	5	12	2
LapSim (manufacturer: Surgical-Science, Sweden) ⁴⁵	Virtual reality (TS, MIS): time, motion, errors	13	383	2	1	0	2	2	10	2
GI Mentor I/II (manufacturer: Symbionix Corp., USA) ⁴⁶	Virtual reality (TS, endoscopy): time, motion, errors	12	383	2	2	0	1	1	11	0
VIST (Vascular Intervention System Training; manufacturer: Mentice AB, Sweden) ⁴⁷	Virtual reality (TS, MIS): time, contrast used, success, complications	11	400	2	0	0	2	0	9	0
MIST-VR (Minimally Invasive Surgery Trainer-Virtual Reality; manufacturer: Mentice AB, Sweden) ⁴⁸	Virtual reality (TS, MIS): time, motion, errors	10	554	10	0	0	2	2	7	1
ProMIS (manufacturer: Haptica, Ireland) ⁴⁹	Augmented reality box trainer (TS, MIS): time, motion	10	500	6	1	0	3	1	8	1
EyeSi (manufacturer: VR Magic, Germany) ⁵⁰	Virtual reality (TS, ophthalmologic surgery): time, errors	7	181	0	1	0	0	0	7	0
LapMentor (manufacturer: Symbionix Corp., USA) ⁵¹	Virtual reality (TS, MIS): time, time-accuracy	7	302	4	0	0	1	0	5	0
Xitact LS500 (manufacturer: Xitact SA, Switzerland) ⁵²	Virtual reality (TS, MIS): time, motion, errors	5	554	0	0	0	1	1	4	0
IC-SAD (Imperial College Surgical Assessment Device) ⁵³	Motion tracking system (TS, generic): time, motion	3	101	12	1	0	0	2	3	0

*TS indicates technical skills; MIS, minimally invasive surgery; NTS, nontechnical skills.

†No. of studies in which the tool was selected for primary coding (see Method for selection process). Validity evidence is coded only for primary studies.

‡No. of studies in which the tool was considered secondary.

§No. of studies in which this evidence source was presented.

Table 5

Methodological Quality of Studies Included in a 2011 Systematic Review of Technology-Enhanced Simulation for Assessment

Scale item	Response (weighting value)	No. (%) present	
		All studies, N = 417	Studies reviewed in detail, N = 217
Medical Education Research Study Quality Instrument (MERSQI)*			
Study design (maximum weighting: 3)	One-group single assessment (1)	225 (54)	119 (55)
	One-group pre–post (1.5)	152 (36)	79 (36)
	Observational two-group (2)	8 (2)	2 (1)
	Randomized two-group (3)	32 (8)	17 (8)
Sampling: no. of institutions (maximum weighting: 1.5)	1 (0.5)	340 (82)	174 (80)
	2 (1)	39 (9)	18 (8)
	>2 (1.5)	38 (9)	25 (12)
Sampling: follow-up (maximum weighting: 1.5)	<50% or not reported (0.5)	117 (28)	49 (23)
	50%–74% (1)	6 (1)	3 (1)
	≥75% (1.5)	294 (71)	165 (76)
Type of data: outcome assessment (maximum weighting: 3)	Cannot tell (0)	2 (1)	0
	Subjective (1)	17 (4)	2 (1)
	Objective (3)	398 (95)	215 (99)
Validity evidence (maximum weighting: 3)	Content (1)	214 (51)	166 (77)
	Internal structure (1)	179 (43)	162 (75)
	Relations with other variables (1)	367 (88)	201 (93)
Data analysis: appropriate (maximum weighting: 1)	Appropriate (1)	378 (91)	198 (91)
Data analysis: sophistication (maximum weighting: 2)	Descriptive (1)	18 (4)	6 (3)
	Beyond descriptive analysis (2)	399 (96)	211 (97)
Highest outcome type (maximum weighting: 3)	No quantitative outcome	1 (0.2)	0
	Satisfaction, attitudes, perceptions (1)	13 (3)	0
	Knowledge, skills (1.5)	368 (88)	196 (90)
	Behaviors (2)	30 (7)	19 (9)
	Patient/health care outcomes (3)	5 (1)	2 (1)
Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2)[†]			
Participant selection	Low risk of bias	—	25 (12)
Index test: conduct or interpretation	Low risk of bias	—	85 (39)
Index test: match with target condition	Low concern about applicability	—	205 (94)
Reference test: conduct or interpretation (N = 102)	Low risk of bias	—	34 (33)
Reference test: match with target condition (N = 102)	Low concern about applicability	—	83 (81)
Flow of participants	Low risk of bias	—	153 (71)

*The MERSQI was coded for all 417 studies. Numbers in parentheses indicate weighting as per original MERSQI description.²⁶ Mean overall score was 12.3 (SD 1.8), and median score was 12.5 (range, 4 to 17).

[†]The QUADAS-2²⁵ was coded for 217 studies reporting two or more elements of validity evidence. The QUADAS-2 also has a question about the appropriateness (applicability to the review question) of participants, but this was not coded because it was an exclusion criterion for the review itself.

with 58 (27%) studies reporting eligibility criteria, 38 (18%) reporting any sampling method, and 30 (14%) reporting the number eligible. Twelve studies (6%) failed to report the number enrolled, and 52 (24%) failed to define the training level for all trainees.

Only 19 (9%) studies reported sample size calculations. Among 198 studies

correlating two variables or comparing two groups, at least one main statistical method was undefined in 8 (4%). Similarly, among 153 studies reporting reliability analyses, statistical methods were undefined in 18 (12%). Confidence intervals were reported for 2% of correlation coefficients (N = 2 of 92 studies) and 10% of reliability coefficients (N = 16 of 153 studies).

Discussion and Conclusions

Brennan^{34(p8)} stated, “Validity theory is rich, but the practice of validation is often impoverished.” This systematic review of simulation-based assessment suggests that such is unfortunately the case in this field of medical education. Most of the 417 studies in this sample offered only limited validity evidence, and nearly half reported only one element of new

evidence. By far the most commonly reported source of validity evidence—and the sole source for one-third of studies—was the relatively weak design of expert–novice comparison. The average number of validity elements decreased slightly or remained constant in more recent studies, suggesting that conditions are not improving. Fewer than two-thirds of the studies proposed an outline of the validity evidence they expected to accrue, and one-fifth failed to interpret the results of the evidence presented. Only six studies acknowledged the current unified evidence-oriented framework.²⁰

We also evaluated methodological quality using the MERSQI and QUADAS-2. Whereas MERSQI overall scores are somewhat higher than those reported in previous studies,^{8,23,26,35} QUADAS-2 ratings indicate a high prevalence of selective inclusion (case–control studies), incomplete description of the population, and lack of rater blinding—all of which have been associated with bias in clinical research.^{36,37} If such associations hold true in education, the findings of such studies may differ from the true properties of the assessment activity.

Reporting quality as appraised using the STARD and GRRAS criteria was also limited. The STARD guidelines were established to ensure reporting of key study features required to appraise the risk of bias (e.g., the information needed to complete instruments such as the QUADAS-2). Indeed, we often found it difficult or impossible to appraise methodological rigor (i.e., using the QUADAS-2 and MERSQI) when key information was reported vaguely or not at all.

Limitations and strengths

This review has limitations. We did not attempt to determine the direction or strength of validity evidence or judge the validity of interpretations for individual tools. However, by focusing on the type of validity evidence reported, we were able to comment on strengths and weaknesses across a diverse field and to prepare a catalog of tools for assessing many different skills (see Table 4, Supplemental Digital Table 1, <http://links.lww.com/ACADMED/A130>). Although beyond the scope of the present work, further evaluation of validity evidence for specific skills domains and for studies reporting

on multiple tools would provide additional insight into these issues.

We found modest interrater agreement for some variables, due at least in part to incomplete or unclear reporting. We addressed this by reaching consensus on all reported data. The low interrater agreement for the QUADAS-2 applicability questions suggests that this instrument may require further refinement or clarification prior to widespread use in education.

We cannot exclude the possibility of publication bias, particularly in those studies exploring associations with clinical outcomes.

We included studies regardless of study design or validity evidence presented. However, we abstracted information for only one assessment tool per report, and we applied the QUADAS-2, STARD, and GRRAS criteria to only half the articles. Although we presume that studies reporting less validity evidence would fare less favorably on these measures, we did not confirm this, and the quality of reporting and methods remains unknown for studies not selected for detailed review.

Comparison with previous reviews

The present review agrees with previous reviews of assessment in simulation-based^{9–16} and non-simulation-based^{26,38–41} education in concluding that validation research is generally lacking. The present review builds on previous reviews by applying a modern validity framework to the field of simulation-based assessment and providing a detailed summary of validity evidence for currently available tools.

Previous reviews of reporting and methodological quality in medical education have focused on studies of educational interventions^{23,26,42} and identified significant shortcomings therein. We are not aware of any study applying the QUADAS-2, STARD, or GRRAS guidelines to medical education. The data we present regarding reporting and methodological quality thus constitute a unique contribution.

Implications for research

Our findings suggest that current validation research methods could

lead to biased results. Education researchers can minimize potential bias by avoiding selective inclusion (i.e., including studies that selectively enroll experienced and inexperienced participants), describing the number and characteristics of the eligible population, and blinding raters to trainee experience, other raters, and (when present) the results of any tests used as related measures.

Although this study focused on simulation-based assessment, we suspect that there is room for improvement in the completeness and clarity of reporting for assessment research in health professions education generally. As noted previously, “Rote adherence to guidelines will not compensate for poor-quality research or inferior writing skills, but inclusion of the elements listed in guidelines ... will enable a wide range of consumers to understand and apply the study results.”²³ It may be useful to further refine the STARD and GRRAS for widespread application to educational assessment research. In the meantime, researchers might use these instruments (or our operational adaptation) to facilitate complete reporting.

Yet guidelines alone will be inadequate, in part because many authors are unaware of them or lack the skills to apply them in practice. True improvement in reporting quality will require the active efforts of journal editors and reviewers who understand, endorse, and enforce relevant standards.²²

Implications for practice

This catalog of tools, indexed by clinical topic, will be of great practical value to educators searching for evidence-based assessment instruments. Unfortunately, only a handful of these tools have been subjected to validation across different assessment contexts. For most of the tools listed in Table 4, we see a preponderance of evidence for relations with other variables (especially relations with training status), and a relative lack of evidence for the content, internal structure, response process, and consequences of scores. It seems that a tool’s widespread use often outstrips the accumulation of validity evidence. To resolve this, researchers must do more than employ robust research methods. They will also need to deliberately target

key evidence sources. This, in turn, would benefit from a structured agenda or argument, as has been proposed for professionalism.⁴³ In the meantime, educators should ensure that actions based on an assessment's scores are commensurate with the strength of the available evidence.

It is often said that assessment drives learning, and as medical education evolves toward personalized training and competency-based decisions,⁵ the role of educational assessment will only enlarge.⁴⁴ Assessments that rely on self-report, log books, hours of training, or written tests to determine procedural competence will no longer suffice. To this end, we need both innovative tools—many of which will involve simulation—and coherent validity arguments supporting the interpretation of scores. Validity arguments, in turn, require rigorous, well-reported research providing strategically collected evidence. An arsenal of tools thus validated will enable decisions regarding formative feedback, mastery of technical and nontechnical skills, remediation, and credentialing that will streamline training and ensure the quality of health professionals and patient care.

Acknowledgments: The authors thank Jason Szostek, MD, Amy Wang, MD, and Patricia Erwin, MLS, for their efforts in article identification, and Colin West, MD, PhD, for his critical review of the manuscript (all affiliated with Mayo Clinic College of Medicine, Rochester, Minnesota).

Funding/Support: No external funding. This work was supported by an award from the Division of General Internal Medicine, Mayo Clinic.

Other Disclosures: None.

Ethical approval: Not applicable.

Previous presentations: An abstract based on this work was presented at the 2012 Simulation Summit of the Royal College of Physicians and Surgeons of Canada, Ottawa, Ontario, Canada, November 2012.

References

- Antiel RM, Thompson SM, Reed DA, et al. ACGME duty-hour recommendations—A national survey of residency program directors. *N Engl J Med*. 2010;363:e12.
- West CP, Tan AD, Habermann TM, Sloan JA, Shanafelt TD. Association of resident fatigue and distress with perceived medical errors. *JAMA*. 2009;302:1294–1300.
- Albanese M, Mejicano G, Gruppen L. Perspective: Competency-based medical education: A defense against the four horsemen of the medical education apocalypse. *Acad Med*. 2008;83:1132–1139.
- Emanuel EJ, Fuchs VR. Shortening medical training by 30%. *JAMA*. 2012;307:1143–1144.
- Weinberger SE, Pereira AG, Iobst WF, Mechaber AJ, Bronze MS; Alliance for Academic Internal Medicine Education Redesign Task Force II. Competency-based education and training in internal medicine. *Ann Intern Med*. 2010;153:751–756.
- Albanese MA, Mejicano G, Mullan P, Kokotailo P, Gruppen L. Defining characteristics of educational competencies. *Med Educ*. 2008;42:248–255.
- Ziv A, Wolpe PR, Small SD, Glick S. Simulation-based medical education: An ethical imperative. *Acad Med*. 2003;78:783–788.
- Cook DA, Hatala R, Brydges R, et al. Technology-enhanced simulation for health professions education: A systematic review and meta-analysis. *JAMA*. 2011;306:978–988.
- Kardong-Edgren S, Adamson KA, Fitzgerald C. A review of currently published evaluation instruments for human patient simulation. *Clin Simul Nursing*. 2010;6:e25–e35.
- Van Nortwick SS, Lendvay TS, Jensen AR, Wright AS, Horvath KD, Kim S. Methodologies for establishing validity in surgical simulation studies. *Surgery*. 2010;147:622–630.
- Ahmed K, Jawad M, Abboudi M, et al. Effectiveness of procedural simulation in urology: A systematic review. *J Urol*. 2011;186:26–34.
- Feldman LS, Sherman V, Fried GM. Using simulators to assess laparoscopic competence: Ready for widespread use? *Surgery*. 2004;135:28–42.
- Schout BM, Hendriks AJ, Scheele F, Bemelmans BL, Scherpbier AJ. Validation and implementation of surgical simulators: A critical review of present, past, and future. *Surg Endosc*. 2010;24:536–546.
- Byrne AJ, Greaves JD. Assessment instruments used during anaesthetic simulation: Review of published studies. *Br J Anaesth*. 2001;86:445–450.
- Edler AA, Fanning RG, Chen MI, et al. Patient simulation: A literary synthesis of assessment tools in anesthesiology. *J Educ Eval Health Prof*. 2009;6:3.
- Fitzgerald TN, Duffy AJ, Bell RL, Berman L, Longo WE, Roberts KE. Computer-based endoscopy simulation: Emerging roles in teaching and professional skills assessment. *J Surg Educ*. 2008;65:229–235.
- Downing SM. Validity: On meaningful interpretation of assessment data. *Med Educ*. 2003;37:830–837.
- Kane MT. Validation. In: Brennan RL, ed. *Educational Measurement*. 4th ed. Westport, Conn: Praeger; 2006:17–64.
- Messick S. Validity. In: Linn RL, ed. *Educational Measurement*. 3rd Ed. New York, NY: American Council on Education and Macmillan; 1989:13–103.
- American Educational Research Association; American Psychological Association; National Council on Measurement in Education. *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association; 1999.
- Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: Theory and application. *Am J Med*. 2006;119:166.e7–166.16.
- Cook DA, Beckman TJ, Bordage G. Quality of reporting of experimental studies in medical education: A systematic review. *Med Educ*. 2007;41:737–745.
- Cook DA, Levinson AJ, Garside S. Method and reporting quality in health professions education research: A systematic review. *Med Educ*. 2011;45:227–238.
- Price EG, Beach MC, Gary TL, et al. A systematic review of the methodological rigor of studies evaluating cultural competence training of health professionals. *Acad Med*. 2005;80:578–586.
- Whiting PF, Rutjes AW, Westwood ME, et al; QUADAS-2 Group. QUADAS-2: A revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med*. 2011;155:529–536.
- Reed DA, Cook DA, Beckman TJ, Levine RB, Kern DE, Wright SM. Association between funding and quality of published medical education research. *JAMA*. 2007;298:1002–1009.
- Bossuyt PM, Reitsma JB, Bruns DE, et al; Standards for Reporting of Diagnostic Accuracy. Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD initiative. *Ann Intern Med*. 2003;138:40–44.
- Kottner J, Audigé L, Brorson S, et al. Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. *J Clin Epidemiol*. 2011;64:96–106.
- Moher D, Liberati A, Tetzlaff J, Altman DG; PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *Ann Intern Med*. 2009;151:264–269, W64.
- Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33:159–174.
- 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.
- Derossis AM, Fried GM, Abrahamowicz M, Sigman HH, Barkun JS, Meakins JL. Development of a model for training and evaluation of laparoscopic skills. *Am J Surg*. 1998;175:482–487.
- American Psychological Association. *Standards for Educational and Psychological Tests and Manuals*. Washington, DC: American Psychological Association; 1966.
- Brennan RL. Perspectives on the evolution and future of educational measurement. In: Brennan RL, ed. *Educational Measurement*. 4th ed. Westport, Conn: Praeger; 2006:1–16.
- Reed DA, Beckman TJ, Wright SM, Levine RB, Kern DE, Cook DA. Predictive validity evidence for medical education research study quality instrument scores: Quality of submissions to JGIM's Medical Education Special Issue. *J Gen Intern Med*. 2008;23:903–907.
- Lijmer JG, Mol BW, Heisterkamp S, et al. Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA*. 1999;282:1061–1066.
- Whiting P, Rutjes AW, Reitsma JB, Glas AS, Bossuyt PM, Kleijnen J. Sources of variation and bias in studies of diagnostic accuracy: A systematic review. *Ann Intern Med*. 2004;140:189–202.

- 38 Beckman TJ, Cook DA, Mandrekar JN. What is the validity evidence for assessments of clinical teaching? *J Gen Intern Med*. 2005;20:1159–1164.
- 39 Kogan JR, Holmboe ES, Hauer KE. Tools for direct observation and assessment of clinical skills of medical trainees: A systematic review. *JAMA*. 2009;302:1316–1326.
- 40 Howley L, Szauter K, Perkowski L, Clifton M, McNaughton N; Association of Standardized Patient Educators (ASPE). Quality of standardised patient research reports in the medical education literature: Review and recommendations. *Med Educ*. 2008;42:350–358.
- 41 Ratanawongsa N, Thomas PA, Marinopoulos SS, et al. The reported validity and reliability of methods for evaluating continuing medical education: A systematic review. *Acad Med*. 2008;83:274–283.
- 42 Baernstein A, Liss HK, Carney PA, Elmore JG. Trends in study methods used in undergraduate medical education research, 1969–2007. *JAMA*. 2007;298:1038–1045.
- 43 Clauser BE, Margolis MJ, Holtman MC, Katsufakis PJ, Hawkins RE. Validity considerations in the assessment of professionalism. *Adv Health Sci Educ Theory Pract*. 2012;17:165–181.
- 44 Holmboe ES, Sherbino J, Long DM, Swing SR, Frank JR. The role of assessment in competency-based medical education. *Med Teach*. 2010;32:676–682.
- 45 Duffy AJ, Hogle NJ, McCarthy H, et al. Construct validity for the LAPSIM laparoscopic surgical simulator. *Surg Endosc*. 2005;19:401–405.
- 46 Ritter EM, McClusky DA 3rd, Lederman AB, Gallagher AG, Smith CD. Objective psychomotor skills assessment of experienced and novice flexible endoscopists with a virtual reality simulator. *J Gastrointest Surg*. 2003;7:871–877.
- 47 Hsu JH, Younan D, Pandalai S, et al. Use of computer simulation for determining endovascular skill levels in a carotid stenting model. *J Vasc Surg*. 2004;40:1118–1125.
- 48 Grantcharov TP, Rosenberg J, Pahle E, Funch-Jensen P. Virtual reality computer simulation. *Surg Endosc*. 2001;15:242–244.
- 49 Botden SM, Buzink SN, Schijven MP, Jakimowicz JJ. ProMIS augmented reality training of laparoscopic procedures face validity. *Simul Healthc*. 2008;3:97–102.
- 50 Rossi JV, Verma D, Fujii GY, et al. Virtual vitreoretinal surgical simulator as a training tool. *Retina (Philadelphia, Pa)*. 2004;24:231–236.
- 51 Zhang A, Hünnerbein M, Dai Y, Schlag PM, Beller S. Construct validity testing of a laparoscopic surgery simulator (Lap Mentor): Evaluation of surgical skill with a virtual laparoscopic training simulator. *Surg Endosc*. 2008;22:1440–1444.
- 52 Rosenthal R, Gantert WA, Scheidegger D, Oertli D. Can skills assessment on a virtual reality trainer predict a surgical trainee's talent in laparoscopic surgery? *Surg Endosc*. 2006;20:1286–1290.
- 53 Datta V, Mackay S, Mandalia M, Darzi A. 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.