

What counts as validity evidence? Examples and prevalence in a systematic review of simulation-based assessment

David A. Cook · Benjamin Zendejas · Stanley J. Hamstra ·
Rose Hatala · Ryan Brydges

Received: 12 October 2012 / Accepted: 9 April 2013 / Published online: 2 May 2013
© Springer Science+Business Media Dordrecht 2013

Abstract Ongoing transformations in health professions education underscore the need for valid and reliable assessment. The current standard for assessment validation requires evidence from five sources: content, response process, internal structure, relations with other variables, and consequences. However, researchers remain uncertain regarding the types of data that contribute to each evidence source. We sought to enumerate the validity evidence sources and supporting data elements for assessments using technology-enhanced simulation. We conducted a systematic literature search including MEDLINE, ERIC, and Scopus through May 2011. We included original research that evaluated the validity of simulation-based assessment scores using two or more evidence sources. Working in duplicate, we abstracted

Electronic supplementary material The online version of this article (doi:[10.1007/s10459-013-9458-4](https://doi.org/10.1007/s10459-013-9458-4)) contains supplementary material, which is available to authorized users.

D. A. Cook
Office of Education Research, Mayo Medical School, Rochester, MN, USA

D. A. Cook (✉)
Division of General Internal Medicine, Mayo Clinic College of Medicine, Mayo 17-W, 200 First Street
SW, Rochester, MN 55905, USA
e-mail: cook.david33@mayo.edu

B. Zendejas
Department of Surgery, Mayo Clinic College of Medicine, Rochester, MN, USA
e-mail: zendejas.benjamin@mayo.edu

S. J. Hamstra
Academy for Innovation in Medical Education, Faculty of Medicine, University of Ottawa, Ottawa,
ON, Canada
e-mail: shamstra@uottawa.ca

R. Hatala
Department of Medicine, University of British Columbia, Vancouver, BC, Canada
e-mail: rhatala@mac.com

R. Brydges
Department of Medicine, University of Toronto, Toronto, ON, Canada
e-mail: ryan.brydges@utoronto.ca

information on the prevalence of each evidence source and the underlying data elements. Among 217 eligible studies only six (3 %) referenced the five-source framework, and 51 (24 %) made no reference to any validity framework. The most common evidence sources and data elements were: relations with other variables (94 % of studies; reported most often as variation in simulator scores across training levels), internal structure (76 %; supported by reliability data or item analysis), and content (63 %; reported as expert panels or modification of existing instruments). Evidence of response process and consequences were each present in <10 % of studies. We conclude that relations with training level appear to be overrepresented in this field, while evidence of consequences and response process are infrequently reported. Validation science will be improved as educators use established frameworks to collect and interpret evidence from the full spectrum of possible sources and elements.

Keywords Medical education · Simulation · Validity · Reliability · Assessment · Evaluation · Reporting quality · Methods quantitative · Educational technology

Introduction

Ongoing transformations in the training of health professionals underscore the need for valid and reliable assessment. (Irby et al. 2010; Holmboe et al. 2010; Naik et al. 2012) The CanMEDS Physician Competency Framework and the Accreditation Council for Graduate Medical Education (ACGME) core competencies are but two examples of the growing emphasis on explicit standards of performance. Trainee assessment plays a pivotal role in decisions regarding program admissions, advancement, remediation, certification, and lifelong learning. As competency-based training systems evolve they will increasingly rely upon performance assessments that support defensible and reproducible decisions. (Holmboe et al. 2010) This bespeaks the need for a robust enterprise to establish the validity of such decisions and the scores that inform them (Schuwirth and van der Vleuten 2011; Boulet et al. 2011).

Historically, validity in the classical framework was divided into three types: content validity, criterion validity (with its subdivisions of concurrent and predictive validity), and construct validity (American Psychological Association 1966). However, over 20 years ago Messick proposed an alternative framework (Messick 1989), and the American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME) adopted this as a standard for the field in 1999. (American Educational Research Association et al. 1999) In this unified model all validity is considered construct validity, and evidence is collected from five sources, namely content, response process, internal structure, relations with other variables, and consequences (defined below). This evidence can then be compiled into a coherent validity argument used to support or refute the proposed score interpretations and uses (Kane 2006; Messick 1989).

However, in health professions education uncertainty remains regarding what types of data could be collected and reported within each evidence classification. For example, a researcher might ask, “What data should I collect to support the content evidence of my new instrument’s scores?” Or a teacher might ask, “How do the published data for this instrument support the decisions I make based on student scores?” While the data that *should* be collected for a given instrument cannot be prescribed without knowing about the instrument and its intended application, workers in this field would benefit from a menu of options that *could* be considered. We are not aware of previous research attempting to offer such a guide.

To address this gap, we sought to enumerate the validity evidence reported for assessments in a focused field of health professions education, namely technology-enhanced simulation. We recently completed a comprehensive systematic review of assessment in technology-enhanced simulation, identifying over 400 articles for review Cook et al. (2013). As part of that review, we collected detailed information on the validity evidence reported in a purposeful sub-sample of approximately half of these studies. Our purpose is to use these studies of simulation-based assessment as a convenience sample to illustrate how validity evidence might be operationally defined and collected for each of the five sources noted above, not only for simulation-based assessments but for educational assessments broadly.

Conceptual definitions

Detailed definitions of the five sources of validity evidence can be found elsewhere, (American Educational Research Association et al. 1999; Cook and Beckman 2006; Downing 2003) and we will provide specific examples below. However, these sources may be briefly defined conceptually as follows:

- *Content evidence* comprises a description of steps taken to ensure that assessment content (including scenarios, questions, response options, and instructions) reflects the construct it is intended to measure (e.g., “professionalism”). This might involve basing the assessment on prior instruments, obtaining expert review, or using an assessment blueprint.
- *Response process evidence* comprises theoretical and empirical analyses evaluating how well rater or examinee actions (responses) align with the intended construct. This includes assessment security (those who cheat are not responding based on the intended construct), quality control, and analysis of examinees’ or raters’ thoughts or actions during the assessment activity.
- *Internal structure evidence* comprises data evaluating the relations among individual assessment items and how these relate to the overarching construct. This most often takes the form of measures of reproducibility (reliability) across items, stations, or raters, but can also include item analysis (item difficulty and item discrimination) and factor analysis.
- *Relations with other variables evidence* regards the statistical associations between assessment scores and another measure or feature that has a specified theoretical relationship. This relationship might be strongly positive (e.g., two measures that should measure the same construct) or negligible (for measures that should be independent).
- *Consequences evidence* regards the impact, beneficial or harmful, of the assessment itself and the decisions and actions that result (e.g. remediation following sub-standard performance). This also includes factors that directly influence the rigor of such decisions, such as the definition of the passing score (e.g., at what point is remediation required?) and differences in scores among subgroups where performance ought to be similar (suggesting that decisions may be spurious).

Methods

This review was planned, conducted, and reported in adherence to PRISMA standards of quality for reporting systematic reviews (Moher et al. 2009). We conducted this review of

simulation-based *assessment* (Cook et al. 2013) concurrently with a review of simulation-based *training* (Cook et al. 2011). Although conducted concurrently, these reviews addressed different questions and employed distinct inclusion criteria. The present report presents previously unreported data from the review of assessment; we summarize only briefly the methods reported previously and focus on those unique to the present report.

Questions

We sought to answer the questions: within the five sources of validity evidence, what specific data elements have been collected to support the validity of scores from technology-enhanced simulation-based assessments, and how has this evidence been used to construct a validity argument?

Evaluating the validity of education assessments

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 Tables 1, 2, 3, 4, 5 for definitions). We also coded the specific elements constituting data within each evidence source. We initially developed a list of elements through a careful examination of previous conceptual expositions (Messick 1989; American Educational Research Association et al. 1999; Downing 2003; Kane 2006; Cook and Beckman 2006)

Table 1 Validity evidence data element definitions and prevalence: content evidence

Data element	Data element definition	Prevalence no. (%) N = 217	Rater agreement (ICC) ^a
Any		137 (63)	0.71
Expert panel	Group consensus or expert review, including formal task analysis	79 (36)	0.69
Previously validated instrument	New instrument based on (or modified from) a previously validated instrument	40 (18)	0.48
Pilot testing and revision	Iterative instrument development involving both pilot test and revision	34 (16)	0.52
Clinical guidelines	Use of clinical guidelines or similar high quality evidence to determine key elements	28 (13)	0.53
Scoring framework	Use of theory, framework, or test blueprint to develop the scoring method	25 (12)	0.41
Test blueprint	Use of theory, framework, or test blueprint to develop instrument items	20 (9)	0.31
Content-construct mismatch ^b	Serious flaw in alignment between construct and test (test content irrelevant to construct, or construct under-represented)	11 (5)	0.30

The prevalence of content evidence here is lower than that for the MERSQI data in Appendix Table 1 of ESM because we required new evidence (as compared with citing evidence previously published)

^a ICC values 0.21–0.4 are considered “fair,” 0.41–0.6 “moderate,” 0.61–0.8 “substantial,” and >0.8 “almost perfect”

^b Negative evidence (i.e., evidence of invalidity), although authors often suggested revisions to improve for future use

Table 2 Validity evidence data element definitions and prevalence: internal structure evidence

Data element	Data element definition	Prevalence no. (%) N = 217	Rater agreement (ICC)
Any		165 (76)	0.84
Reliability, any	Reproducibility of scores across any facet of variation	153 (71)	0.90
... Inter-rater reliability	Reproducibility across different raters	117 (54)	0.91
... Internal consistency	Reproducibility across different items on the test	46 (21)	
... Inter-station reliability	Reproducibility across different stations or tasks	40 (18)	0.71
... Test–retest reliability	Reproducibility across different versions of the test	19 (9)	0.84
... Other reliability	Reproducibility across another facet of variation	1 (0.5)	–
Critical analysis of data distribution	Evaluation of data distribution, missing responses, and/or outliers in terms of appropriateness of scoring	21 (10)	0.26
Empiric scoring	Analysis of empiric data to support scoring method	13 (6)	0.34
Item analysis	Evaluation of inter-item correlation or item discrimination	12 (5.5)	0.49
Factor analysis	Exploratory or confirmatory factor analysis	2 (1)	0.98

See Table 1 for note regarding ICC interpretation

and reviews, (Beckman et al. 2005) and then extended and refined this list as we inductively identified additional data elements.

We classified the validity framework referenced by the authors as reflecting a classical model (American Psychological Association 1966) (i.e., content, criterion, and construct validity), the currently-standard five sources of evidence model, (American Educational Research Association et al. 1999) Kane’s validation model (Kane 2006) (scoring, generalization, extrapolation, and decision), or another model. Finally, we evaluated the completeness of the validity argument using a simplified version of the framework proposed by Kane: (1) statement of the proposed interpretation, (2) linking observed performance to an interpretation (encompassing inferences of scoring, generalization, and extrapolation), and (3) linking the interpretation to a decision.

Study eligibility

We included studies published in any language that had as a stated purpose the evaluation of a tool for assessing health professions learners, at any stage in training or practice, using technology-enhanced simulation. While the original review (Cook et al. 2013) made no restrictions based on validity evidence reported, the present study selected only studies reporting 2 or more sources of validity evidence. We focused on relatively robust studies in order to enrich our sample, recognizing that our estimates will suggest a higher prevalence of validity evidence than is true for the field generally. Interrater agreement for this full/abbreviated coding decision was excellent (intraclass correlation coefficient [ICC], 0.80). 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” (Cook et al. 2011).

Table 3 Validity evidence data element definitions and prevalence: relations with other variables evidence

Data element	Data element definition	Prevalence no. (%) N = 217	Rater agreement (ICC)
Any		204 (94)	0.82
Learner characteristic, any	Association with a hypothetically influential trainee characteristic	168 (77)	0.81
... Learner characteristic: general training	Association with general level of training (example: experienced surgeon vs first-year resident)	147 (68)	0.85
... Learner characteristic: task- specific training	Association with training status for this specific task, participants otherwise at same stage of training (example: fourth-year medical students, trained vs. untrained as part of this study)	33 (13)	0.69
... Learner characteristic: other	Other characteristics: scripted level of performance, self-reported confidence, video game experience	3 (1)	0.24
Separate measure, any	Association with a separate measure (no restriction to time)	102 (47)	
Separate measure, concurrent	Association with a separate measure within 1 month of initial assessment	98 (45)	0.86
... patient care	Association with a measure of performance in clinical practice with actual patients	11 (5)	0.85
... different simulator	Association with a measure of performance on a different simulator or standardized patient	15 (7)	0.87
... same simulator	Association with a measure of performance using the same simulator (i.e., same simulator, different metric)	60 (28)	0.80
... computer, written, or oral exam	Association with a measure of performance by computer, written, or oral exam	23 (11)	0.80
Separate measure, delayed (predictive)	Association with a separate measure ≥ 1 month apart	7 (3)	0.66
... patient care	Association with a measure of performance in clinical practice with actual patients	6 (3)	0.60
... different simulator	Association with a measure of performance on a different simulator or standardized patient	1 (0.5)	–
... computer, written, or oral exam	Association with a measure of performance by computer, written, or oral exam	1 (0.5)	–

See Table 1 for note regarding ICC interpretation

Study identification

Our search strategy has been previously published. (Cook et al. 2011) We searched multiple databases, including MEDLINE, ERIC, and Scopus, for relevant articles using a search strategy developed by an experienced research librarian. We used no beginning date cutoff, and the last date of search was May 11, 2011.

Study selection

Working in duplicate, reviewers screened all candidate studies for inclusion. We first reviewed each title and abstract, then if needed we reviewed the full text of studies judged

Table 4 Validity evidence data element definitions and prevalence: response process evidence

Data element	Data element definition	Prevalence no. (%) N = 217	Rater agreement (ICC)
Any		13 (6)	0.35
Rater data analysis ^a	Analysis of rater disagreements or errors to determine the underlying cause	5 (2.5)	0.22
Test security	Problems with, or empirically-determined strengths of, systems to ensure the integrity and security of data after collection	4 (2)	0.40
Problems with video capture ^a	Problems with the collection of video (for example, suboptimal camera angle)	2 (1)	0.22
Effect of rater training	Evaluation of the impact of rater training ^b	1 (0.5)	–
Rationale for composite outcomes	Justification for the combination of disparate data to form a single overall score (as per Messick 1989)	1 (0.5)	–
Think aloud protocol ^c	Analysis of raters' verbal description of thought processes while rating	0	–

See Table 1 for note regarding ICC interpretation

^a Negative evidence (i.e., evidence of invalidity), although authors often suggested revisions to improve for future use

^b We required evidence showing that the training achieved desired endpoints. The mere presence of rater training did not count

^c We expected to find this element based on its presence in conceptual discussions regarding this evidence source, but found no instances

eligible or uncertain. We resolved conflicts by consensus. Interrater agreement for study inclusion was substantial (ICC = 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 validity evidence sources, elements, and frameworks as outlined above, resolving conflicts by consensus. Most studies employed >1 assessment tool, and for these we selected the tool with the strongest validity evidence. We also evaluated study methods using the Medical Education Research Study Quality Instrument (MERSQI) (Reed et al. 2007).

We summarized the data using counts and, where appropriate, means. For validity screening interrater agreement ICC's ranged 0.67–0.91 except for response process (ICC = 0.34, raw agreement 95 %) and consequences (ICC = 0.56). Per Landis and Koch, (Landis and Koch 1977) interrater agreement values 0.21–0.4 are considered “fair,” 0.41–0.6 “moderate,” 0.61–0.8 “substantial”, and >0.8 “almost perfect.” Interrater agreement for individual elements is reported in Tables 1, 2, 3, 4 and 5. We used SAS 9.3 for all analyses.

For each evidence source we selected prototypical quotes from included articles to illustrate how this evidence might be reported.

Table 5 Validity evidence data element definitions and prevalence: consequences evidence

Data element	Data element definition	Prevalence no. (%) N = 217	Rater agreement (ICC)
Any		20 (9)	0.65
Rigorous pass/fail cutpoint, established approach	Pass/fail cutpoint determined using an established standard setting procedure (these included: Angoff method, modified borderline group method, Markov modeling, ROC curve [with citation], and nonlinear causal resource analysis)	6 (3)	0.61
Rigorous pass/fail cutpoint, unproven approach	Pass/fail cutpoint determined using a method described by author without reference to prior approaches (these included: experienced surgeons as the benchmark, natural breaks in the data distribution, and ROC curve [no citation])	8 (4)	0.58
Evaluation of actual pass rate	Comparison of actual vs expected number of passes/failures	3 (1.5)	0.10
Anticipated impact	Anticipated impact (positive, negative, or neutral) of testing on students or patients	2 (1)	0.44
Differential item functioning	Actual consistencies or inconsistencies in test performance across learner groups, in comparison with what would be expected for these groups	1 (0.5)	–
Unanticipated impact ^a	Unanticipated impact (positive, negative, or neutral) of testing on students or patients	0	–

See Table 1 for note regarding ICC interpretation

^a We expected to find this element based on its presence in conceptual discussions regarding this evidence source, but found no instances

Results

Trial flow and participants

From 10,911 potentially relevant articles, we identified 417 studies evaluating technology-enhanced assessment (Appendix Figure 1 of ESM). Nineteen studies reported no validity evidence at all, despite having as a stated goal the evaluation of an assessment instrument, and 181 reported only one validity evidence source. These 200 studies were excluded from the present review. Thus, the final sample consisted of 217 studies enrolling 11,519 trainees (median 32 trainees per study, interquartile range 20–60). One article was published in Spanish, the remainder in English. Table 6 summarizes study characteristics, including information on the trainees, topics, and simulator outcome measures.

Study quality

Appendix Table 1 of ESM summarizes study quality as evaluated using the MERSQI. Notably, 43 of the studies involved more than one institution and all but two reported objective measures of performance. However, nearly one-fourth lost more than half the participants before follow-up or failed to report follow-up.

Table 6 Description of studies

Study characteristic	Level	No. studies	No. participants ^a	
All studies		217	11,519	
Participants ^b	Medical students	55	2,032	
	Physicians in postgraduate training	148	4,290	
	Physicians in practice	92	1,342	
	Nurses and nursing students	20	1,242	
	Emergency medical technicians and students	2	83	
	Dentists and dental students	4	200	
	Other	10	62	
	Ambiguous/mixed	52	2,268	
	Clinical topics ^{b,c}	Minimally invasive surgery	63	2,809
		Other surgery	51	2,841
Endoscopy and ureteroscopy		21	580	
Resuscitation/trauma training		35	2,587	
Nontechnical skills ^d		27	2,042	
Anesthesia		28	1,764	
Endovascular procedures		6	171	
Physical examination		7	1,236	
Obstetrics		6	253	
Airway management		5	281	
Simulator outcomes ^b	Skill: time	59	2,649	
	Skill: process	203	10,926	
	Skill: product	21	1,089	

See Appendix Table 1 of ESM for details on individual studies

^a Numbers reflect the number enrolled. Most studies included trainees from >1 level

^b The number of studies and learners sum to more than the number for all studies because most studies included >1 learner group or reported multiple outcomes, and several fit within >1 clinical topic

^c Selected listing of the topics addressed most often (several other topics were addressed, with lower frequency)

^d Nontechnical skills include communication, leadership/team management, organization, situational awareness, and decision making

Validity framework and argument

Seventy-five studies (34 %) cited use of the “classical” validity framework (American Psychological Association 1966) (i.e., content, criterion, and construct validity) for planning and interpreting their data. Seventy-six (35 %) cited use of a more limited framework, such as one component of the classical framework (construct, content, or concurrent validity alone [in 48, 8, and 6 studies, respectively]), or a validity framework not accepted as part of the classical view (e.g., “discriminant validity”). Of note, when used in this context the term “construct validity” nearly always referred specifically to data showing differences between trainee training levels (as compared with the more encompassing perspective espoused by Messick (1989) and adopted by the AERA/APA/NCME

Table 7 Validity interpretive argument: components and prevalence

Argument component	No. (%) present, N = 217	
	Proposed ^a	Final ^a
Proposed interpretation (construct)	187 (86)	–
How evidence will support the link between observations and interpretation, all	139 (64)	168 (77)
Evidence of trainee discrimination	100 (46)	130 (60)
Any other evidence	74 (34)	106 (49)
How score interpretations could be used to make decisions	11 (5)	41 (19)

^a The proposed interpretive argument appeared before presentation of data/results. The final interpretive argument appeared after data/results

(American Educational Research Association et al. 1999). Nine studies (4 %) reported “reliability” as their only evidentiary framework. Six (3 %) invoked the currently-accepted five sources of validity evidence model (American Educational Research Association et al. 1999). None mentioned Kane’s validation model. Fifty-one studies (24 %) did not mention any framework.

Kane proposed that a validity interpretive argument requires at least three elements: the proposed interpretation, the link between observed performance and interpretation, and the link between interpretations and the decision or use to which these interpretations will be applied. (Kane 2006) As shown in Table 7, 187 studies (86 %) stated the proposed interpretation or construct (such as “laparoscopic surgery skill” or “team leadership skill”). We coded separately the *proposed* argument (an a priori statement of how to interpret the data that will be collected) and the *final* argument (an interpretation of the data observed). For both the proposed and final argument, most studies discussed the link between observations and interpretations, but very few noted the link between score interpretations and the decisions that could be justified based on those interpretations (see Table 7 for details). In alignment with the prevalence of validity evidence reported above, interpretive arguments involving trainee discrimination (e.g. “showing a difference between trainee levels will support the validity of scores”) were more common than all other arguments put together.

Validity evidence

Tables 1, 2, 3, 4 and 5 summarize the prevalence of validity evidence sources presented in the 217 articles, along with detailed information regarding the data elements that comprised that evidence. By far the most common evidence source was relations with other variables (N = 204 studies), followed by internal structure (N = 165) and content (N = 137). Fewer than 10 % of studies reported evidence of consequences (N = 20) or response process (N = 13).

The single most common validity evidence element was the analysis of how simulator scores varied according to a learner characteristic such as training status (procedural experience or training level; N = 168 studies). The association between simulator scores and another concurrently-measured variable (e.g., scores on another simulation assessment or a written exam) was also frequently reported (N = 98). Both of these contribute evidence of relations with other variables. One illustrative example is:

The residents obtained higher [technical skills global rating scores] than did the medical students. ... Separate analyses of variance were conducted on each of the 5 items of the communication form. There was a main effect of level of training for the coherence item. However, there were no main effects of level of training for empathy, verbal communication, and nonverbal communication items. ... Performance scores on the 2 assessments of the technical skills - the checklist and the GRS - correlated strongly with each other overall ($r = 0.760$). ... However, performance on the communication scale was not correlated significantly with [technical skills] scores ($r = 0.06$). ... These results suggest that technical and communication skills vary independently of each other. (LeBlanc et al. 2009) (*associations with training level and with two measures that should correlate together, and absence of correlation with measures that should be independent*)

Content evidence provides information about how instrument items or scoring rubrics were developed. Here, the most common data elements were expert panels ($N = 79$), modification of a previously validated instrument ($N = 40$), and pilot testing with subsequent revision ($N = 34$). Unexpected findings from other evidences sources (e.g. poor inter-rater reliability or response process flaws) were occasionally ($N = 11$) traced back to problems with instrument content (e.g., item wording or scoring criteria).

A search of the literature was conducted to locate and review existing teamwork measurement instruments. ... These instruments were reviewed to identify applicable categories, elements and items. This exercise resulted in a list of fifteen teamwork elements and 57 items which were subsequently collapsed by an accredited resuscitation expert into a list of 27 items. Members of the research team further refined the number of elements to eight. The evidence indicates that the binary nature of checklists tend to overlook the more holistic components of clinical competence with the suggestion that global ratings of performance are appropriate. The next phase of design required expert clinicians to rate the content of the instrument using a five-point scale on the relevance of each item. (Cooper et al. 2010) (*prior literature as foundation, expert review, and justification for scale*)

Based on a blueprint representing the curricular goals and objectives for undergraduate students in emergency medicine a ten station OSCE was designed. ... All scenarios were piloted and evaluated amongst students as well as amongst emergency physicians and paramedics to refine the scenarios, to revise the checklists and to define the time frame for the stations. Furthermore, the whole circle was evaluated in test runs with faculty members as examinees. (Ruesseler et al. 2010) (*test blueprint and pilot testing with revision*)

For six items with a low kappa coefficient, there was either an obvious difference in interpretation of the criteria for a correct score, the criteria were not exactly followed or there were no clear criteria. (van der Heide et al. 2006) (*use of other evidence to identify content-related flaws*)

Internal structure evidence evaluates relationships among assessment items. Interrater reliability was often reported ($N = 117$) among internal structure evidence, while internal consistency ($N = 46$) and inter-station ($N = 40$) reliability were less common. Factor analysis was rarely reported ($N = 2$). Generalizability theory, a unifying approach to reliability that estimates the variability of multiple facets (e.g., stations, raters, and items)

simultaneously, Bloch and Norman (2012) was used to estimate reproducibility for 15 studies.

A principal component factor analysis using Varimax rotation was conducted based on a resuscitation expert's ratings of 56 previously video-recorded resuscitation team events. ... To establish the internal consistency/reliability of the eleven individual TEAM items a Cronbach alpha coefficient was calculated. ... We made a random selection of six of the 56 (11 %) resuscitation video-recordings to determine inter-rater reliability by examining the level of agreement between two experts on the scale scores. Cohen's kappa for the 11 item scale produced an interrater reliability of 0.55 which is considered by Fliess to be 'fair'. ... Test-retest reliability was explored through blinded re-rating after an interval of six months, by the same expert. ... Cohen's kappa for the eleven item scale produced a test-retest reliability of 0.53. (Cooper et al. 2010) (*factor analysis, internal consistency, inter-rater reliability, and test-retest reliability*)

The following summary measures were produced. First, item difficulty (p) - the proportion of participants who received credit for the item. ... Comparisons of item p values (and their means) can be used investigate performance differences between the intern and the management team. ... The second measure was item discrimination (D) - the correlation between the item-level score and the total checklist score. Here, higher values (i.e., $D > 0.30$) indicate that the item is able to discriminate between low- and high-ability individuals/teams. (Ottestad et al. 2007) (*item analysis [item difficulty and discrimination]*)

Overall reliability of the participant scores [was] determined using the generalizability theory. This analysis is used to estimate and quantify the sources of variability in the participant scores, including those connected with the raters, the specific scenarios, the participants, and the associated interactions. ... The generalizability coefficient (reliability estimate) for the simulation scores, based on eight scenarios and two raters, was 0.80. The main sources of variance in participant scores could be attributed to differences in scenario difficulty and to the resultant fluctuations in individual performance from one encounter to the next (task sampling variability). (Henrichs et al. 2009) (*generalizability theory*)

Response process evidence evaluates how well rater or examinee responses align with the intended construct. Only 13 studies reported evidence from this source. The most common element contributing to response process evidence was the analysis of rater raw data to explain low reliability or other rater errors. Six studies noted problems with data integrity such as test security or problems with video capture. Although several studies reported that rater training was *done*, this alone does not support score validity (since the training might be inefficient); only one study reported the *impact* of rater training. (Rosenstock et al. 2004)

For 11 items the kappa coefficient was below 0.50. ... The item assessing the evaluation of the response to volume expansion [was] difficult to see on the videotape because the evaluation took place at the end of the scenario. Three items were not scored by some raters, probably due to doubts about what to score as the performance could not be seen properly on the videotape. The kappa coefficient for the oxygen administration item and the assessment of the appropriateness of the decision to intubate was low, because the criteria for when to score or not apparently were not

clear to all raters. (van der Heide et al. 2006) (*analysis of rater errors to identify problems with response process*)

The interobserver agreement was satisfactory between observers trained in CRM rating ($r = 0.43$), whereas the observer without education in CRM assessment gave significantly lower CRM scores compared to trained observers ($r = 0.09$ and $r = -0.01$, respectively). (Rosenstock et al. 2004) (*impact of rater training*)

Evidence of consequences looks at the impact of the assessment and the resultant decisions, and factors that directly impact those decisions. Only 20 studies reported this source of evidence. Among these, the most common element of consequences evidence was the method of establishing a pass/fail cutpoint (pass/fail standard) (Downing 2003)—using either established methods ($N = 5$) or a description of a user-created standard-setting approach ($N = 9$). There were other standard-setting approaches that we felt did *not* count as validity evidence, including “by group consensus” with no additional details of the process ($N = 3$), adopting a previously reported standard with no verification in the new setting ($N = 3$), and using a standard with no mention of its derivation ($N = 13$). Only two studies explored an anticipated impact on students or patients, (Berkenstadt et al. 2006; Stefanidis et al. 2009) and none reported an unanticipated impact. Three studies contrasted the number of actual vs expected passing grades to determine whether the test was unfair and/or whether the pass/fail cutpoint was appropriate. (Hesselfeldt et al. 2005; Hatala et al. 2008; Hemman et al. 2007) One study noted a difference between two learner groups on some subscales but not others (differential item functioning), and considered score invalidity as a possible explanation before concluding that the difference was most likely real (LeBlanc et al. 2009).

[Methods] At the end of each task repetition, the simulator provides feedback on task duration and motion tracking metrics that consist of pathlength and smoothness. ... Participants practiced on the object positioning task until expert-derived proficiency levels for time, path length, and smoothness were achieved. ... Motion metrics were considered valuable if the training duration was extended based on proficiency attainment in all metrics compared with [number of repetitions] alone. ... [Results] Four participants benefited from the motion metrics as their training was prolonged by an average of 25 repetitions. (Stefanidis et al. 2009) (*impact of the assessment activity*)

Once we agreed on criteria for the CVC checklist, we employed the Angoff method to establish MPSs [minimum passing scores]. ... We distributed the CVC checklist to a panel of eight pulmonary critical care or anesthesia critical care experts from five institutions. [followed by a brief description of Angoff method as implemented] (Huang et al. 2009) (*establishment of a passing standard*)

The residents only outperformed the medical students on the coherence subscale of the communication assessment [but not] on other subscales of empathy, verbal communication, and nonverbal communication. There are several potential explanations for this absence of difference. ... First, it is possible that the IPPI format does not allow for the discrimination of communication skills, or that the communication scale we used is not sensitive enough to detect differences in communication. Second, it is possible that our raters may have shown little ability to discriminate between different levels of communication abilities. Alternatively, it is possible that the residents' communication skills are not superior to those of medical students.

Previous researchers have shown that [coherence may be related to training but that] empathy and nonverbal communication skills may be more constant traits of the individuals. ... As such, these findings, ... suggest that residents most likely do not have generally superior communication skills than the fourth year medical students. (LeBlanc et al. 2009) (*consideration of differential item functioning*)

Discussion

To help educators better understand the evidence sources and the constituent data elements informing the validity of assessments in health professions education, we conducted a comprehensive review of research in technology-enhanced simulation. We found that nearly all studies in this sample reported evidence of relations with other variables, and about two-thirds reported evidence of internal structure and content. By contrast, very few studies reported evidence to support the response processes or consequences of the assessment scores.

More than just the broad categories of evidence, however, we also report a breakdown of the specific data elements that comprised this evidence, and specific examples for many of these elements. For example, we found that content evidence can be supported by at least seven different elements of data (see Table 1) and that internal structure can be supported by at least four data elements in addition to reliability (Table 2). Most evidence exploring relations with another variable used training level as the covariate, while fewer than half used a separately-measured concurrent or predictive variable. We identified six data elements each for response process (Table 4) and consequences (Table 5), although each of these was reported rarely, if at all. We trust these operational definitions and examples will enable educators to seek and recognize evidence from various sources.

We also found that most investigators use an outdated or incomplete framework to interpret validity data, if they use any framework at all. Only six studies invoked the model adopted as a field standard in 1999, (American Educational Research Association et al. 1999) and none used the newer model recently proposed by Kane (2006). While most authors clearly defined the intended construct, only about two-thirds prospectively defined how evidence would support the link between observed performance and desired inferences. Fewer than 20 % explained how score interpretations could support defensible decisions.

Limitations

We had suboptimal interrater agreement for response process evidence generally and for some specific elements of validity evidence, especially those reported less often. This could be due to ambiguous reporting (for example, authors rarely identified evidence of consequences or response process as such), imprecise definitions, or insufficient examples for rater training. Regardless, this highlights the need for greater clarity in the definitions for these elements. As regards the prevalence estimates, we reached consensus on all reported data.

We did not attempt to determine the direction or strength of validity evidence, or judge the validity of interpretations for individual tools. We also did not attempt to collect information about the specific components comprising Kane's validation model, (Kane 2006) as we expected (and confirmed) this model to be infrequently used. While Kane's framework is insightful, useful, (Hawkins et al. 2010; Schuwirth and van der Vleuten

2012) and of potentially high impact, the AERA/APA/NCME framework remains the current standard.

We restricted inclusion to studies reporting two or more sources of evidence, and as such the prevalence data reflect a biased sample (skewed toward higher quality) compared with the field as a whole. Also, for studies that reported more than one assessment instrument we abstracted information only for the instrument with strongest validity evidence. Since our primary intent was to understand the type of data elements generally, rather than to evaluate specific instruments, these limitations likely had little bearing on our conclusions.

Comparison with previous reviews

Similar to Beckman et al. (2005) we have attempted to identify generalizable principles in the field of validation science by examining a focused area of medical education. Our present work expands upon the work of those authors in at least three substantial ways: our sample is ten-fold larger; we collected detailed information on the elements contributing to each evidence source; and we coded the validity framework and validity interpretive argument.

Previous reviews of assessment instruments (Beckman et al. 2005; Ratanawongsa et al. 2008; Kogan et al. 2009) found content evidence in 34–55 % of studies, internal structure evidence in 40–74 %, and relations with other variables evidence in 5–64 %. By contrast, far fewer studies report evidence regarding response process and consequences; one review found no studies (of 22) reporting strong evidence from either source, (Beckman et al. 2005) and another review found only one study (of 55) reporting response process other than rater training (Kogan et al. 2009). Reviews of assessment in simulation-based education (Van Nortwick et al. 2010; Edler et al. 2009; Kardong-Edgren et al. 2010; Ahmed et al. 2011; Feldman et al. 2004; Cook et al. 2013) have likewise concluded that validation research is generally lacking in this field.

Implications for practice and research

This work advances the science of validation by extending our understanding of what constitutes validity evidence. By clarifying the different types of data that can contribute evidence to the validity argument, researchers will be better able to plan studies, and both researchers and practitioners will be better able to interpret results and identify gaps in existing evidence. We trust this will be particularly useful for evidence of consequences and response process, which have been notoriously difficult for people to understand and apply.

One problem with the current five-evidence-sources model is the absence of prescription regarding which evidence sources to cite for a given intended interpretation or decision. While our efforts provide granularity in the data elements that comprise each evidence source, this work is not prescriptive. Kane's model (Kane 2006) offers a step forward in this regard by explicitly outlining the steps of a generic validity argument. In future work, it may be useful to map the data elements identified in our study onto Kane's framework, thus bridging these two frameworks.

We caution that the relative proportions of validity evidence or the underlying data elements should not be interpreted as prescriptive of best practices. On the contrary, we believe that in many cases the prevalence of reported evidence reflects issues of convenience or use of a narrow view of validity evidence. For example, the most common data

element was the association between test scores and the learner's training level, also known as trainee discrimination. Also, among the studies that proposed an interpretive argument, trainee discrimination was the most common rationale given to link observations and inferences. Yet trainee discrimination contributes only weak validity evidence. Discrimination may be useful if it is absent (i.e., scores that do *not* discriminate experts from novices are likely invalid). However, the presence of trainee discrimination contributes minimally to the validity of score interpretations, because a difference in scores could very plausibly be due to expert-novice differences other than those relevant to the intended construct. As Norman poignantly noted, by this standard gray hair and baldness would be good measures of expertise when comparing senior surgeons and third-year medical students (Geoff Norman, personal communication).

We recommend greater emphasis on describing and defending the *uses* of scores, and especially the decisions made and the actions that follow such decisions. Validity in general, and consequences evidence in particular, is most meaningful when evaluated in the context of a specific decision and action (e.g., certify/remediate or accept/reject). The consequences of such actions constitute the most important evidence, in the same way that the ultimate usefulness of a clinical diagnostic test is determined by its beneficial and harmful impacts on patients. (Hubbard et al. 2011; Lam et al. 2010; Woolf and Harris 2012)

Finally, we suspect that many authors may not have known about the five evidence sources model, and we cannot expect that authors thus unaware would collect and report validity evidence that they did not know was relevant let alone important. Yet this implies that validation, as currently represented in the literature, is a rather haphazard process. Our findings appear to reinforce Cronbach's lament, now 25 years old: "The weak program [of validation without a clear framework and hypothesis] is sheer exploratory empiricism. ... The strong program ... calls for making one's theoretical ideas as explicit as possible, then devising deliberate challenges." (Cronbach 1988, pp. 12–13) As with other frameworks, models, and theories, (Bordage 2009) validity frameworks guide the deliberate collection and meaningful interpretation of research data. Perhaps the most important implication of our work for future practice and research is to enhance awareness that validity frameworks exist, and that investigators—both those intending to disseminate their work, and those evaluating instruments for local use—should follow a rigorous approach to validation that outlines an a priori argument targeting the weakest assumptions in the evidentiary chain.

Conclusions

Our intent was to illustrate how validity evidence might be operationally defined and collected. We encourage researchers and practitioners to seek a variety of validity evidence elements and sources when conducting or appraising validation research, and interpret this evidence using a formal validity framework.

Acknowledgments The authors thank Jason Szostek, MD, Amy Wang, MD, and Patricia Erwin, MLS, for their efforts in article identification. This work was supported by an award from the Division of General Internal Medicine, Mayo Clinic.

Conflict of interest None.

No human subjects involved None.

References

- Ahmed, K., Jawad, M., Abboudi, M., Gavazzi, A., Darzi, A., Athanasiou, T., et al. (2011). Effectiveness of procedural simulation in urology: A systematic review. *Journal of Urology*, *186*, 26–34.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- American Psychological Association. (1966). *Standards for Educational and Psychological Tests and Manuals*. Washington, DC: American Psychological Association.
- Beckman, T. J., Cook, D. A., & Mandrekar, J. N. (2005). What is the validity evidence for assessments of clinical teaching? *Journal of General Internal Medicine*, *20*, 1159–1164.
- Berkenstadt, H., Ziv, A., Gafni, N., & Sidi, A. (2006). The validation process of incorporating simulation-based accreditation into the anesthesiology Israeli national board exams. *Israel Medical Association Journal: IMAJ*, *8*, 728–733.
- Bloch, R., & Norman, G. (2012). Generalizability theory for the perplexed: A practical introduction and guide: AMEE Guide No. 68. *Medical Teacher*, *34*, 960–992.
- Bordage, G. (2009). Conceptual frameworks to illuminate and magnify. *Medical Education*, *43*, 312–319.
- Boulet, J. R., Jeffries, P. R., Hatala, R. A., Korndorffer, J. R., Jr, Feinstein, D. M., & Roche, J. P. (2011). Research regarding methods of assessing learning outcomes. *Simulation in Healthcare*, *6*(Suppl), S48–S51.
- Cook, D. A., & Beckman, T. J. (2006). Current concepts in validity and reliability for psychometric instruments: Theory and application. *American Journal of Medicine*, *119*, 166.e7–166.e16.
- Cook, D. A., Brydges, R., Zendejas, B., Hamstra, S. J., & Hatala, R. (2013). Technology-enhanced simulation to assess health professionals: A systematic review of validity evidence, research methods, and reporting quality. *Academic Medicine*. doi:10.1097/ACM.0b013e31828ffdfc.
- Cook, D. A., Hatala, R., Brydges, R., Zendejas, B., Szostek, J. H., Wang, A. T., et al. (2011). Technology-enhanced simulation for health professions education: A systematic review and meta-analysis. *JAMA*, *306*, 978–988.
- Cooper, S., Cant, R., Porter, J., Sellick, K., Somers, G., Kinsman, L., et al. (2010). Rating medical emergency teamwork performance: Development of the Team Emergency Assessment Measure (TEAM). *Resuscitation*, *81*, 446–452.
- Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 3–17). Hillsdale, NJ: Routledge.
- Downing, S. M. (2003). Validity: On the meaningful interpretation of assessment data. *Medical Education*, *37*, 830–837.
- Edler, A. A., Fanning, R. G., Chen, M. I., Claire, R., Almazan, D., Struyk, B., et al. (2009). Patient simulation: A literary synthesis of assessment tools in anesthesiology. *Journal of Education, Evaluation and Health Profession*, *6*, 3.
- Feldman, L. S., Sherman, V., & Fried, G. M. (2004). Using simulators to assess laparoscopic competence: Ready for widespread use? *Surgery*, *135*, 28–42.
- Hatala, R., Issenberg, S. B., Kassen, B., Cole, G., Bacchus, C. M., & Scalese, R. J. (2008). Assessing cardiac physical examination skills using simulation technology and real patients: A comparison study. *Medical Education*, *42*, 628–636.
- Hawkins, R. E., Margolis, M. J., Durning, S. J., & Norcini, J. J. (2010). Constructing a validity argument for the mini-Clinical Evaluation Exercise: A review of the research. *Academic Medicine*, *85*, 1453–1461.
- Hemman, E. A., Gillingham, D., Allison, N., & Adams, R. (2007). Evaluation of a combat medic skills validation test. *Military Medicine*, *172*, 843–851.
- Henrichs, B. M., Avidan, M. S., Murray, D. J., Boulet, J. R., Kras, J., Krause, B., et al. (2009). Performance of certified registered nurse anesthetists and anesthesiologists in a simulation-based skills assessment. *Anesthesia and Analgesia*, *108*, 255–262.
- Hesselfeldt, R., Kristensen, M. S., & Rasmussen, L. S. (2005). Evaluation of the airway of the SimMan™ full-scale patient simulator. *Acta Anaesthesiologica Scandinavica*, *49*, 1339–1345.
- Holmboe, E. S., Sherbino, J., Long, D. M., Swing, S. R., & Frank, J. R. (2010). The role of assessment in competency-based medical education. *Medical Teacher*, *32*, 676–682.
- Huang, G. C., Newman, L. R., Schwartzstein, R. M., Clardy, P. F., Feller-Kopman, D., Irish, J. T., et al. (2009). Procedural competence in internal medicine residents: Validity of a central venous catheter insertion assessment instrument. *Academic Medicine*, *84*, 1127–1134.
- Hubbard, R. A., Kerlikowske, K., Flowers, C. I., Yankaskas, B. C., Zhu, W., & Miglioretti, D. L. (2011). Cumulative probability of false-positive recall or biopsy recommendation after 10 years of screening mammography: A cohort study. *Annals of Internal Medicine*, *155*, 481–492.

- Irby, D. M., Cooke, M., & O'Brien, B. C. (2010). Calls for reform of medical education by the Carnegie Foundation for the Advancement of Teaching: 1910 and 2010. *Academic Medicine*, *85*(2), 220–227.
- Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport: Praeger.
- Kardong-Edgren, S., Adamson, K. A., & Fitzgerald, C. (2010). A review of currently published evaluation instruments for human patient simulation. *Clinical Simulation in Nursing*, *6*, e25–e35.
- Kogan, J. R., Holmboe, E. S., & Hauer, K. E. (2009). Tools for direct observation and assessment of clinical skills of medical trainees: A systematic review. *JAMA*, *302*, 1316–1326.
- Lam, L. L., Cameron, P. A., Schneider, H. G., Abramson, M. J., Muller, C., & Krum, H. (2010). Meta-analysis: Effect of B-type natriuretic peptide testing on clinical outcomes in patients with acute dyspnea in the emergency setting. *Annals of Internal Medicine*, *153*(11), 728–735.
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*, 159–174.
- LeBlanc, V. R., Tabak, D., Kneebone, R., Nestel, D., MacRae, H., & Moulton, C.-A. (2009). Psychometric properties of an integrated assessment of technical and communication skills. *American Journal of Surgery*, *197*, 96–101.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: American Council on Education and Macmillan.
- Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *Annals of Internal Medicine*, *151*, 264–269.
- Naik, V. N., Wong, A. K., & Hamstra, S. J. (2012). Review article: Leading the future: Guiding two predominant paradigm shifts in medical education through scholarship. *Canadian Journal of Anaesthesia*, *59*, 213–223.
- Ottestad, E., Boulet, J. R., & Lighthall, G. K. (2007). Evaluating the management of septic shock using patient simulation. *Critical Care Medicine*, *35*, 769–775.
- Ratanawongsa, N., Thomas, P. A., Marinopoulos, S. S., Dorman, T., Wilson, L. M., Ashar, B. H., et al. (2008). The reported validity and reliability of methods for evaluating continuing medical education: A systematic review. *Academic Medicine*, *83*, 274–283.
- Reed, D. A., Cook, D. A., Beckman, T. J., Levine, R. B., Kern, D. E., & Wright, S. M. (2007). Association between funding and quality of published medical education research. *JAMA*, *298*, 1002–1009.
- Rosenstock, C., Ostergaard, D., Kristensen, M. S., Lippert, A., Ruhnau, B., & Rasmussen, L. S. (2004). Residents lack knowledge and practical skills in handling the difficult airway. *Acta Anaesthesiologica Scandinavica*, *48*, 1014–1018.
- Ruesseler, M., Weinlich, M., Byhahn, C., Muller, M. P., Junger, J., Marzi, I., et al. (2010). Increased authenticity in practical assessment using emergency case OSCE stations. *Advances in Health Sciences Education*, *15*, 81–95.
- Schuwirth, L. W., & van der Vleuten, C. P. (2011). General overview of the theories used in assessment: AMEE Guide No. 57. *Medical Teacher*, *33*(10), 783–797.
- Schuwirth, L. W., & van der Vleuten, C. P. (2012). Programmatic assessment and Kane's validity perspective. *Medical Education*, *46*, 38–48.
- Stefanidis, D., Scott, D. J., & Korndorffer, J. R., Jr. (2009). Do metrics matter? Time versus motion tracking for performance assessment of proficiency-based laparoscopic skills training. *Simulation in Healthcare: Journal of the Society for Simulation in Healthcare*, *4*, 104–108.
- van der Heide, P. A., van Toledo-Eppinga, L., van der Heide, M., & van der Lee, J. H. (2006). Assessment of neonatal resuscitation skills: A reliable and valid scoring system. *Resuscitation*, *71*, 212–221.
- Van Nortwick, S. S., Lendvay, T. S., Jensen, A. R., Wright, A. S., Horvath, K. D., & Kim, S. (2010). Methodologies for establishing validity in surgical simulation studies. *Surgery*, *147*, 622–630.
- Woolf, S. H., & Harris, R. (2012). The harms of screening: New attention to an old concern. *JAMA*, *307*, 565–566.