Consequences Validity Evidence: Evaluating the Impact of Educational Assessments

David A. Cook, MD, MHPE, and Matthew Lineberry, PhD

Abstract

Because tests that do not alter management (i.e., influence decisions and actions) should not be performed, data on the consequences of assessment constitute a critical source of validity evidence. Consequences validity evidence is challenging for many educators to understand, perhaps because it has no counterpart in the older framework of content, criterion, and construct validity. The authors' purpose is to explain consequences validity evidence and propose a framework for organizing its collection and interpretation.

Both clinical and educational assessments can be viewed as interventions. The act of administering or taking a test, the interpretation of scores, and the ensuing decisions and actions influence those being assessed (e.g., patients or students) and other people and systems (e.g., physicians, teachers, hospitals, schools). Consequences validity evidence examines such impacts of assessments. Despite its importance, consequences evidence is reported infrequently in health professions education (range 5%–20% of studies in recent systematic reviews) and is typically limited in scope and rigor.

Consequences validity evidence can derive from evaluations of the impact on examinees, educators, schools, or the end target of practice (e.g., patients or health care systems); and the downstream impact of classifications (e.g., different score cut points and labels). Impact can result from the uses of scores or from the assessment activity itself, and can be intended or unintended and beneficial or harmful. Both quantitative and qualitative research methods are useful. The type, quantity, and rigor of consequences evidence required will vary depending on the assessment and the claims for its use.

Emerging reforms in health professions education such as competency-based education, mastery learning, entrustable professional activities, and adaptive learning environments underscore the need for valid assessments of learning outcomes. The currently standard framework for thinking about assessment validity, first proposed by Messick[1] in 1989, defines validity as "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests."[2] Validity can be viewed as a hypothesis about the meaning (interpretations) and application (uses) of test scores. Like any hypothesis, the validity hypothesis can be tested by collecting evidence, which is then summarized in a coherent narrative or validity argument that identifies strengths, weaknesses, and residual gaps (i.e., the degree of support).[3,4] Evidence targeting key assumptions is vital to crafting a strong validity argument.

In this framework, evidence derives from five different sources: content, internal structure, relationships with other variables, response process, and consequences (see Table 1).[5,6] The first three sources map to prior notions of content validity; reliability; and criterion, correlational, and construct validity, respectively,[7] and as such have been readily understood by educators. However, the concepts of response process and consequences have no counterpart in the older framework, and in our experience it has been challenging for educators to understand these concepts and visualize how these might be implemented in practice. Perhaps for these reasons consequences evidence is rarely reported in health professions education research,[6,8] and when reported it tends to be limited in scope.[6] Yet, authors have repeatedly emphasized the critical significance of consequences evidence in presenting a compelling validity argument.[3,5,6,9]

Although the origin of this disparity between what experts request and what investigators report is not fully known, a detailed discussion of consequences evidence would enhance both awareness of the issue and understanding of how to collect needed evidence. The purpose of this article is to explain consequences evidence in easily understood terms and propose a framework for organizing the collection and interpretation of such evidence along with several examples.

In approaching this topic, we first reviewed seminal works on validity in general[1–3,5,6,9–11] and consequences evidence specifically.[12–17] We also reviewed each article in three systematic reviews of validity evidence in health professions education assessments[6,8,18] to identify the frequency and type of consequences evidence presented therein. We then synthesized these theories and exemplars to create a novel framework for planning and organizing consequences evidence, and to propose specific hypothetical examples of how this evidence might be collected in practice.

What Do We Mean by "Consequences"?

Consequences evidence looks at the impact, beneficial or harmful and intended or unintended, of assessment.[2,13] In this sense, assessment can be viewed as an intervention. The act of

D.A. Cook is professor of medicine and medical education, associate director, Mayo Clinic Online Learning, and consultant, Division of General Internal Medicine, Mayo Clinic College of Medicine, Rochester, Minnesota.

M. Lineberry is assistant professor of medical education, Department of Medical Education, and assistant director for research, Graham Clinical Performance Center, University of Illinois at Chicago, Chicago, Illinois.

Correspondence should be addressed to David A. Cook, Division of General Internal Medicine, Mayo Clinic College of Medicine, Mayo 17-W, 200 First Street SW, Rochester, MN 55905; telephone: (507) 266-4156; e-mail: cook.david33@mayo.edu.

Acad Med. 2016;91:785–795.
First published online February 2, 2016
doi: 10.1097/ACM.0000000000001114

Table 1
Five Sources of Validity Evidence[a]

Source of evidence	Definition and examples
Content	Steps taken to ensure that assessment content (including scenarios, questions, response options, and instructions) reflects the construct it is intended to measure.
	Examples: Adapting items from existing instruments, obtaining expert review, using an assessment blueprint.
Response process	Theoretical and empirical analyses evaluating how well rater or examinee responses align with the intended construct; this includes respondent thought processes, response systems, and test security.
	Examples: Analyzing think-aloud protocols, evaluating rater training, testing procedures to prevent cheating.
Internal structure	Data evaluating the relationships among individual assessment items and how these relate to the overarching construct.
	Examples: Measuring reproducibility (reliability), performing item analysis (item difficulty and item discrimination) or factor analysis.
Relationships with other variables	Associations between assessment scores and another measure or feature that has a specified theoretical relationship.
	Examples: Correlating scores from two different instruments, comparing learner groups expected to differ.
Consequences	Data evaluating the impact, beneficial or harmful, of the assessment itself and the decisions and actions that result, and factors that directly influence the rigor of such decisions.
	Examples: See Table 2 and Appendix 1.

[a]These reflect different sources of evidence, not different types of validity.

administering or taking a test, the analysis and interpretation of scores, and the ensuing decisions and actions (such as remediation, feedback, promotion, or board certification) all have direct impacts on those being assessed and on other people and systems (e.g., teachers, patients, schools). These impacts should ideally be evaluated to determine whether actual benefits align with anticipated benefits and outweigh costs and adverse effects.

An analogy with clinical medicine may help to illustrate the concept of assessments as interventions. Mammograms are assessments (diagnostic tests) used to screen for breast cancer. Current evidence suggests that they are less useful in younger women because interpretation is more difficult, that comparison with old films is often required before a judgment can be made, and that false positives are common and subject women to unnecessary biopsies and emotional stress.[19–23] Yet most experts agree that for women aged 50 to 74, annual screening mammograms are beneficial because they substantially reduce the adverse consequences of breast cancer.[24,25] Despite the imperfections of the test and unintended negative consequences of false-positive results, the test has an overall beneficial impact. However, for younger women (for whom the false-positive rate is higher[20]) and for older women (who might die of other causes before they die of breast cancer) screening mammograms should not be automatic according to some guidelines,[24] although this is a matter of controversy.[26] Other clinical examples include the use of brain natriuretic peptide for diagnosing heart failure,[27] flexible sigmoidoscopy for colon cancer screening,[28] and computed tomographic angiography for detection of coronary artery disease[29]—each of which has been evaluated using randomized trials comparing the long-term impact of testing (and its associated clinical decisions) vs no testing. In each case, the act of testing is in fact an intervention with costs, benefits, and potential harms.

Similarly, educational assessments can be viewed as interventions with potential costs, benefits, and harms. For example, a board certification exam might protect patients from incompetent physicians and encourage physicians to study, but might also force competent physicians with poor test-taking skills to engage in needless remediation. This exam has "intervened" in the lives of physicians and patients and led to both beneficial and harmful consequences. To further illustrate, Table 2 cites published studies in which use of educational assessments improved knowledge and skills, altered study behaviors, enhanced faculty rater skills, or led to curricular change.

Stated another way: Consequences evidence does not address the question, "Are we measuring what we think we are measuring?" (a question answered by the other sources of validity evidence). Rather, it addresses, "Does the activity of measuring and the subsequent interpretation and application of scores achieve our desired results with few negative side effects?"

Investigators occasionally confuse consequences as a source of assessment validity evidence (the focus of this article) with other uses of the word "consequences" (e.g., as a general synonym for impact or outcome). For example, studies often evaluate the consequences of training activities (courses, curricula, online modules, or simulation scenarios) using outcomes measured in a test setting or in real clinical practice; such evaluations of *training* interventions are conceptually distinct from studies evaluating consequences evidence to support *assessment* validity. Alternatively, an assessment validation study might evaluate the association between test scores and other concurrent or future measurements of patients, programs, or society (i.e., real-life outcomes or "consequences"). Such associations would inform the validity argument by establishing "relationships with other variables"[2,7] but would not reflect consequences validity evidence (i.e., the analysis focuses on the relationships among scores rather than the consequences of the assessment itself). Of course, there are situations in which measures of impact constitute evidence of assessment consequences (assessments are, after all, interventions), and correlational analyses can provide consequences evidence (see Table 2 and Appendix 1 for examples). What matters is not the study design or statistical analysis but, rather, how the evidence is presented in the validity argument: Consequences evidence establishes the impact of interpretations and uses of assessment scores.

The Importance of Consequences Evidence

Clinicians are often taught not to order a test if it won't improve patient management. The same holds true for

Table 2
Examples of Consequences Evidence in Published Articles

Type of consequences evidence	Quote
Impact on examinee: Topic-specific knowledge and skill (directly resulting from test)	"Each student videotaped [two] outpatient encounters.… The control group neither saw their videotape nor received any feedback.… The [self-critique] students were provided the first recording and instructed to review it alone using a checklist.… The [critique by preceptor] students reviewed their first videotapes with [a] preceptor.… [Three pediatricians] reviewed each student's second videotape and rated performance using a form [that was adapted from other published rating forms].… Students who received preceptors' critiques on their first videotapes performed significantly better on their subsequent interviews and examinations than either students who received no preceptor-guided feedback or students who made only self-guided critiques."[45]
	"The [interns] took a history and performed a physical examination, observed by the faculty member, and were graded on a detailed 56-item CEX [clinical evaluation exercise] form.… After the presentation, evaluators went over the case in detail with the intern, providing feedback. At the end of the same session, the evaluator and the interns completed separate postfeedback forms which asked for a detailed recall of teaching points made by the evaluator.… The teaching points recalled by the interns on their [postfeedback forms] were counted and compared with the points from the CEX and [evaluators' postfeedback forms].… Interns recalled hearing 46% of the points recorded on the [evaluators' postfeedback forms]. They recalled discussing only 27.3% of the points noted on the initial CEX form."[44] (i.e., *failure* to achieve intended consequence that interns would recall evaluator-provided feedback).
Impact on examinee: Learning behaviors (test preparation)	"Formative assessment was integrated into 1 bedside teaching session per week. Students undertook directly observed BFA [bedside formative assessment]. Clinician educators provided feedback at the end of each session.… Study outcomes [included] the impact of BFA on learning behaviors.… More than 2/3 of students reported an increase in preparatory reading for bedside teaching sessions."[47]
Impact on examinee: Learning behaviors (additional training)	"At the end of each task repetition, the [laparoscopy] simulator provides feedback on task duration and motion tracking metrics that consist of path length 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."[34]
Impact on educators: Assessment and feedback skills	"The evaluator rated the student's performance on one or more of the eight domains, checked off the appropriate global rating, and gave verbal and written feedback to the student.… We compared the end-of-clerkship evaluations from the intervention group with those from a historical control group who had completed the clerkship a year earlier.… Differences between the two groups' scores were statistically significant for … [domains including 'received specific feedback on:' a) 'history taking,' b) 'physical examination,' and c) 'assessment/decision making'; results reported in table] with higher means reported for the intervention group."[40] (i.e., suggests that evaluators' provided better feedback after performing repeated assessments).
Impact on others: Curriculum planning	"As a result of the cumulative collection of these data the teaching of clinical problem solving in the CMC [clinical methods course] is being revised and reinforced to include sessions on [list of new topics]. [Other issues raised by students have] resulted in a review of the use of assessments across the school,… a review of teaching of clinical skills across the 5 years of the course, a request for increased faculty support for clinical skills teaching,… [and] a workshop on 'teaching the patient-centred consultation.'"[39]
Impact on defensibility: Establishment of passing standard	"Once we agreed on criteria for the CVC [central venous catheter] checklist, we employed the Angoff method to establish [minimum passing scores].… We distributed the CVC checklist to a panel of eight pulmonary critical care or anesthesia critical care experts from five institutions [brief description of Angoff method as implemented]."[55]
Impact on defensibility: Consideration of differential item functioning	"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 [integrated procedural performance instrument] 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."[38] (finding differences between learner groups on some subscales but not others indicates differential item functioning; ultimately, the authors concluded this is most likely real rather than a source of score invalidity).

educational assessments: If they do not lead to improved learning outcomes sufficient to outweigh costs and potential harms, they should not be used. Messick[1(p85)] argued that "Evaluation of the consequences and side effects of testing is a key aspect of the validation of test use." Kane's[3(p54)] more recent conceptual reframing of validation, which focuses on key inferences in the validity argument, gives similar priority to evidence supporting the consequences of assessment: "Consequences, or outcomes, are the bottom line in evaluating decision procedures. A decision procedure that does not achieve its goals, or does so at too high a cost, is likely to be abandoned, even if it is based on perfectly accurate information." Other authors have also supported the primacy of consequences evidence.[5,6,9]

Just as the ultimate evidence for the value of a diagnostic test is the impact on practice, the ultimate evidence for the value of an educational assessment is the impact on learners, teachers, and the people and systems they influence.[12] Like clinical tests, educational assessments may fail to realize their intended benefits or may have costs or unintended negative consequences that outweigh the benefits.[12,13,17] In such instances one could argue that the rigor of instrument development, the reliability of scores, and

the strength of score correlations with other variables really don't matter. Such concerns underpin many recent criticisms of high-stakes testing as part of the board recertification process.[30] For this reason, we believe that evidence of consequences is ultimately the most important source of validity evidence.

Consequences Evidence in Health Professions Education Research

Consequences evidence is reported only infrequently in health professions education. A systematic review of 22 clinical teaching assessments found only 2 studies (9%) that reported consequences evidence, and in neither case did the original researchers identify the evidence as such.[8] One study found that providing formative feedback to teachers enhanced their teaching scores,[31] whereas the other study found that the assessment raised awareness of effective teaching behaviors.[32] A systematic review of 417 articles examining simulation-based assessment[6] found only 20 studies (5%) reporting consequences evidence. The majority of this evidence comprised establishing a pass/fail cut point (n = 14). Two studies explored an anticipated impact on students or patients,[33,34] 3 contrasted the number of actual vs. expected passing grades,[35–37] and 1 study noted differential item functioning as a possible source of invalidity.[38] No study reported an unanticipated impact. Finally, a systematic review of 55 studies evaluating assessment tools for direct observation[18] found 11 studies (20%) reporting consequences evidence other than satisfaction with the assessment activity. All of these evaluated the impact of assessment, documenting outcomes including curricular changes based on common deficiencies,[39] improved feedback,[40–43] poor recall of feedback provided (i.e., *failure* to achieve intended consequence),[44] improved objectively measured skills,[45,46] and increased test preparation activities.[47] Table 2 contains illustrative quotes from several of these published studies.

A Framework for Evaluating Consequences Evidence

Consequences evidence consists of data on the impact of an assessment on diverse parties: learners, educators, and educational institutions; patients, providers, and health care institutions; and even society at large. Such impact can be beneficial or harmful, and it may be intentional or unintentional.[13] Intentional benefits are probably the easiest to anticipate and measure; unintentional harms may be the most difficult (because they cannot be easily anticipated or explicitly targeted).[48] Experts have also distinguished direct effects of score use (e.g., instructional guidance or advancement decisions) from indirect effects (e.g., influence on student motivation or preparation activities, instructor lesson plans, and public perceptions).[17] However, although these classifications are helpful for categorizing, interpreting, and reporting consequences evidence once it has been collected, they are inadequate for helping investigators to consider broadly the potential sources of consequences evidence when *planning* an assessment validation study. Moreover, the same effect might be considered intended or unintended, beneficial or harmful, and direct or indirect depending on the proposed theory, interpretation, and use of the assessment. For example, an assessment might have *unintended* effects on learners' general orientations toward performing well relative to peers vs. mastering content for its own sake (performance vs. mastery goal orientations[49]). However, promoting stronger mastery goal orientations may be an explicitly *intended* consequence of assessment when adopting a mastery learning curricular model.[50] Similarly, one could imagine educational assessments that lead physicians to be risk averse in *beneficial* ways (e.g., carefully following protocol for central line placement after a central line assessment) or in *detrimental* ways (e.g., practicing "defensive medicine" by ordering unnecessary lab tests after a test of medical knowledge).

Previous authors, including ourselves, have included evaluations of the rigor, appropriateness, and consistency of classification cut points and labels as consequences evidence.[5–7,50] Although such evidence has direct bearing on the implications and decisions arising from the assessment, on careful reflection we believe it might be more correctly labeled *preconsequences* evidence because it affects, rather than results from, the actual consequences of assessment. With this caveat, we continue to agree that such evidence fits most appropriately as consequences evidence in Messick's framework. (As an aside, we note that in Kane's more recent framework such evidence fits squarely under the inference of "implications and decision."[3,9])

In considering how to help investigators prospectively plan the collection of consequences evidence and help consumers identify evidence gaps, we have integrated the above conceptual elements to create a comprehensive framework for systematically prioritizing and organizing consequences evidence. First, evidence can derive from evaluations of the impact *on* examinees, educators, and other stakeholders (e.g., patients), and the impact *of* classifications ("preconsequences," e.g., different cut scores or labels, and accuracy across examinee subgroups). Second, studies can be distinguished as evaluating the impact of test *score use* (similar to the "direct" effects noted above) such as the effectiveness of score-guided remediation or advancement decisions; or the impact of the assessment *activity itself* (independent of scores) such as change in preassessment study behaviors or the effect of test-enhanced learning. To use a clinical example: A woman might get anxious about an upcoming mammogram because she is scared that it might detect cancer (impact of [anticipated] "score" use), or she might be worried about the potential pain or financial cost (impact of the test activity independent of the "score"). Each of these dimensions could include consequences that are intended or unintended, and beneficial or harmful; adding the latter points completes a four-dimensional framework (see Figure 1). Investigators could use this framework to systematically consider the potential consequences of an assessment, prioritize evidence gaps, and select research approaches to fill these gaps. We briefly discuss below how data might be collected to evaluate impact and defensibility, and illustrate this in Appendix 1 with examples spanning all dimensions.

A straightforward approach to evaluate a test's impact—for both a clinical diagnostic test and for an educational assessment—would be to randomize half of the participants to complete the test and the other half to no test,[13,51] and then quantitatively measure relevant anticipated outcomes, or use qualitative methods to observe for anticipated and unanticipated effects. Of course, local needs may make a randomized trial

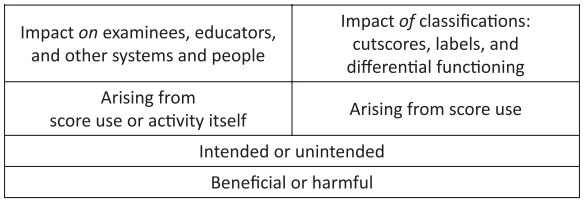

Figure 1 Framework for organizing consequences evidence. Investigators should consider the relevance and priority of each dimension in turn when planning or interpreting consequences evidence. The impact of classifications might be viewed as preconsequences evidence.

infeasible in many situations. A less robust but still useful approach might use less rigorous study designs (such as nonrandomized cohort, single-group pretest–posttest, or even single-group posttest-only studies) but measure the same outcomes. Those being assessed are not the only ones impacted by an assessment. Appendix 1 illustrates potential impacts on educators, patients, institutions, and nonassessed learners.

As noted above, preconsequences evidence includes factors that directly influence the defensibility of classifications based on test results (interpretations and decisions), such as the labels applied to the test itself and any subtests[1]; the definition of the passing score (e.g., at what point is remediation required?)[5]; and differences in scores among subgroups where performance ought to be similar (e.g., men vs. women), suggesting that decisions may be spurious.[52] Finally, investigators could monitor pass/fail rates; for example, a failure rate higher or lower than expected might indicate a test that is either too hard or too easy, respectively.

We distinguish unintended consequences, which can be nonetheless anticipated and prospectively measured, from unforeseeable consequences, which can only be identified after the fact. We further emphasize that data need not be numeric. Qualitative data, properly planned and collected, could provide strong evidence[9]—especially when seeking to identify unintended or unforeseeable consequences.

The data in many of these examples are highly subjective and open to alternative interpretation. For example, score differences among subgroups could be a sign of invalidity if scores should be the same, but could also be interpreted as supporting validity if scores would be expected to vary. Similarly, the ideal failure rate will vary by situation. It is essential to articulate in advance what findings would support or undermine the validity argument,[9,10,53] often guided by a theory of action linking the assessment and its consequences.[3,54] Ultimately, it may be difficult if not impossible to establish a clear cause–effect relationship between the assessment and its consequences.[14] This should not, however, justify educators in ignoring this important element of the validity argument. Triangulation of different evidence sources and data collection methods will help establish a defensible argument.

Finally, the side effects of intended uses of an assessment should not be confused with the effects of misuse.[10] Any application of test scores beyond the scope of existing evidence constitutes, strictly speaking, a misuse. This would include adopting an assessment for new purposes (e.g., using licensure exam scores to inform admissions decisions) or adapting an assessment by changing elements in the instrument, procedures, or learner population. Although it is commonplace and often profitable to adopt or adapt an existing assessment, those doing so should remember that "Test makers are not responsible for negative consequences following from test misuse.… When users appropriate tests for purposes not sanctioned and studied by the test developers, users become responsible for conducting the needed validity investigation."[13(p8)]

Identifying and Using Consequences Evidence in Practice

Not all consequences evidence is equally compelling. Simple improvement in test scores from one testing occasion to the next (e.g., "Students did better when they were retested, suggesting that their skills had improved as a result of the first test") would not, for example, contribute persuasive evidence of consequences because we can imagine plausible alternative explanations for this change (i.e., learning from other experiences). Learner and faculty ratings of satisfaction with the assessment, self-reported improvements in skill attributed to the assessment, and pass/fail rates without a comparison reference point would provide useful but rather weak evidence. Similarly, the establishment of a pass/fail cut point, regardless of how rigorously done, is relatively weak evidence until the consequences of that cut point have been evaluated in practice. Anecdotes without robust quantitative or qualitative data likewise provide only weak support. Stronger evidence will come from studies using a comparison group (randomized, or nonrandomized historical or concurrent control group); objective measures of the desired outcomes that are different from the test itself; or rigorous qualitative data collection and analysis.

Although consequences evidence is the most important source of evidence, test developers, test users, researchers, and journal editors must remember that it constitutes only one of several elements in a comprehensive validity argument. No single source can or should dominate. Moreover, robust consequences "evidence cannot be collected until the test is used as intended for some period of time."[14(p15)] As such, a stepwise approach seems reasonable. We propose that during initial instrument evaluation, developers and researchers might prioritize presumably easier and less costly evidence sources (e.g., content, internal structure/reliability, relationships with other variables, response process [see Table 1]) and then progress to rigorous evaluation of consequences if this evidence proves supportive.

The type, quantity, and rigor of consequences evidence will vary depending on the assessment—more specifically, on the proposed arguments or claims for benefit. For example, a licensure exam that claims to enhance patient safety (anticipated benefit) will impact the employability of physicians who fail. Such an assessment likely merits greater evidence of consequences (e.g., Are anticipated benefits realized? How was the pass/fail cut point established? How often do competent physicians fail?) than an assessment designed to promote feedback to medical students. However, some supposedly "low-stakes"

exams could have potentially significant consequences, especially if implemented on a large scale or repeated over an extended period of time. For example, an assessment intended to promote feedback could have significant cumulative effects across multiple domains of competence, professional identity, self-directed learning, and self-efficacy if administered daily over an entire year of training.

Although our framework is more comprehensive than any other we found, application of this framework will require thoughtful consideration of assessment purposes, procedures, and theory; the level of evidence required; and practical constraints of context.

Unfavorable validity evidence often points to problems elsewhere in the assessment process. Negative consequences can usually be traced back to one of four underlying problems[3]: the measurement or scoring procedure (e.g., irrelevant, unreliable, or omitted test items); the specific interpretation (e.g., an inappropriate pass/fail cut point); the attribute being measured (i.e., the wrong construct); or the response (e.g., the actions that follow the decision). For example, a test intended to identify students in need of remediation in cardiac auscultation might fail to have intended consequences because it contains flawed items, because too many competent students are labeled as incompetent, because it measures knowledge rather than skill, or because the remediation program is ineffective.

Finally, although the present article focuses on education, the importance of assessment consequences is not limited to educational tests. Indeed, the earlier example of the consequences of mammography illustrates the application of this concept to clinical medicine. Other applications would include (but certainly are not limited to) patient symptom scales, teacher rating scales, employment aptitude inventories, customer satisfaction surveys, and research questionnaires.

Concluding Remarks

In conclusion, we emphasize the following. First, assessments are really diagnostic tests, and both in medicine and in education they can be viewed as interventions. Second, consequences validity evidence looks at the impact of assessments (as interventions) on examinees and other stakeholders, and the defensibility of score classifications ("preconsequences" evidence). Such consequences can arise from score use or the assessment activity itself, and can be intentional or unintended and beneficial or harmful. Third, consequences validity evidence is the most important source of evidence because if the assessment does not have the desired impact, it should not be used. Finally, the type, quantity, and rigor of consequences evidence will vary depending on the assessment and the claims for its use.

As health professions educators increasingly rely on assessments to guide important decisions (e.g., recertification, competency-based promotion), they will need stronger evidence to support the validity of the inferences and decisions made. To date, such evidence of consequences has been infrequently reported. Going forward, our framework, which distinguishes the direct impact of the assessment and the indirect influence of other mediating factors and identifies multiple domains within each classification, can help test developers and users to consider a broad view of potentially relevant consequences evidence.

Funding/Support: None reported.

Other disclosures: None reported.

Ethical approval: Reported as not applicable.

References

1 Messick S. Validity. In: Linn RL, ed. Educational Measurement. 3rd ed. New York, NY: American Council on Education and Macmillan; 1989:13–103.
2 American Educational Research Association, American Psychological Association, National Council on Measurement in Education. Validity. In: Standards for Educational and Psychological Testing. Washington, DC: American Educational Research Association; 2014:11–31.
3 Kane MT. Validation. In: Brennan RL, ed. Educational Measurement. 4th ed. Westport, Conn: Praeger; 2006:17–64.
4 Cook DA. When I say… validity. Med Educ. 2014;48:948–949.
5 Downing SM. Validity: On meaningful interpretation of assessment data. Med Educ. 2003;37:830–837.
6 Cook DA, Zendejas B, Hamstra SJ, Hatala R, Brydges R. What counts as validity evidence? Examples and prevalence in a systematic review of simulation-based assessment. Adv Health Sci Educ Theory Pract. 2014;19:233–250.
7 Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: Theory and application. Am J Med. 2006;119:166.e7–166.e16.
8 Beckman TJ, Cook DA, Mandrekar JN. What is the validity evidence for assessments of clinical teaching? J Gen Intern Med. 2005;20:1159–1164.
9 Cook DA, Brydges R, Ginsburg S, Hatala R. A contemporary approach to validity arguments: A practical guide to Kane's framework. Med Educ. 2015;49:560–575.
10 Kane MT. Validating the interpretations and uses of test scores. J Educ Meas. 2013;50:1–73.
11 Kane MT. Validation as a pragmatic, scientific activity. J Educ Meas. 2013;50:115–122.
12 Linn RL. Evaluating the validity of assessments: The consequences of use. Educ Meas Issues Pract. 1997;16:14–16.
13 Shepard LA. The centrality of test use and consequences for test validity. Educ Meas Issues Pract. 1997;16:5–24.
14 Reckase MD. Consequential validity from the test developer's perspective. Educ Meas Issues Pract. 1998;17:13–16.
15 Lane S, Stone CA. Strategies for examining the consequences of assessment and accountability programs. Educ Meas Issues Pract. 2002;21:23–30.
16 Moss PA. Validity in action: Lessons from studies of data use. J Educ Meas. 2013;50:91–98.
17 Haertel E. How is testing supposed to improve schooling? Measurement. 2013;11:1–18.
18 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.
19 Armstrong K, Moye E, Williams S, Berlin JA, Reynolds EE. Screening mammography in women 40 to 49 years of age: A systematic review for the American College of Physicians. Ann Intern Med. 2007;146:516–526.
20 Nelson HD, Tyne K, Naik A, Bougatsos C, Chan BK, Humphrey L; U.S. Preventive Services Task Force. Screening for breast cancer: An update for the U.S. Preventive Services Task Force. Ann Intern Med. 2009;151:727–737, W237.
21 Hubbard RA, Kerlikowske K, Flowers CI, Yankaskas BC, Zhu W, Miglioretti DL. Cumulative probability of false-positive recall or biopsy recommendation after 10 years of screening mammography: A cohort study. Ann Intern Med. 2011;155:481–492.
22 Welch HG, Passow HJ. Quantifying the benefits and harms of screening mammography. JAMA Intern Med. 2014;174:448–454.
23 Roelofs AA, Karssemeijer N, Wedekind N, et al. Importance of comparison of current and prior mammograms in breast cancer screening. Radiology. 2007;242:70–77.
24 U.S. Preventive Services Task Force. Screening for breast cancer: U.S. Preventive Services Task Force recommendation statement. Ann Intern Med. 2009;151:716–726, W-236.
25 American Cancer Society. American Cancer Society recommendations for early breast cancer detection in women without breast symptoms. http://www.cancer.org/cancer/breastcancer/moreinformation/breastcancerearlydetection/breast-cancer-early-detection-acs-recs. Accessed December 15, 2015.

26 Hendrick RE, Helvie MA. United States Preventive Services Task Force screening mammography recommendations: Science ignored. AJR Am J Roentgenol. 2011;196:W112–W116.
27 Lam LL, Cameron PA, Schneider HG, Abramson MJ, Müller C, Krum H. Meta-analysis: Effect of B-type natriuretic peptide testing on clinical outcomes in patients with acute dyspnea in the emergency setting. Ann Intern Med. 2010;153:728–735.
28 Schoen RE, Pinsky PF, Weissfeld JL, et al; PLCO Project Team. Colorectal-cancer incidence and mortality with screening flexible sigmoidoscopy. N Engl J Med. 2012;366:2345–2357.
29 Muhlestein JB, Lappé DL, Lima JA, et al. Effect of screening for coronary artery disease using CT angiography on mortality and cardiac events in high-risk patients with diabetes: The FACTOR-64 randomized clinical trial. JAMA. 2014;312:2234–2243.
30 Teirstein PS. Boarded to death—why maintenance of certification is bad for doctors and patients. N Engl J Med. 2015;372:106–108.
31 Cohen R, MacRae H, Jamieson C. Teaching effectiveness of surgeons. Am J Surg. 1996;171:612–614.
32 Copeland HL, Hewson MG. Developing and testing an instrument to measure the effectiveness of clinical teaching in an academic medical center. Acad Med. 2000;75:161–166.
33 Berkenstadt H, Ziv A, Gafni N, Sidi A. The validation process of incorporating simulation-based accreditation into the anesthesiology Israeli national board exams. Isr Med Assoc J. 2006;8:728–733.
34 Stefanidis D, Scott DJ, Korndorffer JR Jr. Do metrics matter? Time versus motion tracking for performance assessment of proficiency-based laparoscopic skills training. Simul Healthc. 2009;4:104–108.
35 Hesselfeldt R, Kristensen MS, Rasmussen LS. Evaluation of the airway of the SimMan full-scale patient simulator. Acta Anaesthesiol Scand. 2005;49:1339–1345.
36 Hatala R, Issenberg SB, Kassen B, Cole G, Bacchus CM, Scalese RJ. Assessing cardiac physical examination skills using simulation technology and real patients: A comparison study. Med Educ. 2008;42:628–636.
37 Hemman EA, Gillingham D, Allison N, Adams R. Evaluation of a combat medic skills validation test. Mil Med. 2007;172:843–851.
38 LeBlanc VR, Tabak D, Kneebone R, Nestel D, MacRae H, Moulton CA. Psychometric properties of an integrated assessment of technical and communication skills. Am J Surg. 2009;197:96–101.
39 Hastings A, McKinley RK, Fraser RC. Strengths and weaknesses in the consultation skills of senior medical students: Identification, enhancement and curricular change. Med Educ. 2006;40:437–443.
40 Paukert JL, Richards ML, Olney C. An encounter card system for increasing feedback to students. Am J Surg. 2002;183:300–304.
41 Links PS, Colton T, Norman GR. Evaluating a direct observation exercise in a psychiatric clerkship. Med Educ. 1984;18:46–51.
42 Lane JL, Gottlieb RP. Structured clinical observations: A method to teach clinical skills with limited time and financial resources. Pediatrics. 2000;105(4 pt 2):973–977.
43 Ross R. A clinical-performance biopsy instrument. Acad Med. 2002;77:268.
44 Kroboth FJ, Hanusa BH, Parker SC. Didactic value of the clinical evaluation exercise. Missed opportunities. J Gen Intern Med. 1996;11:551–553.
45 Scheidt PC, Lazoritz S, Ebbeling WL, Figelman AR, Moessner HF, Singer JE. Evaluation of system providing feedback to students on videotaped patient encounters. J Med Educ. 1986;61:585–590.
46 Stone H, Angevine M, Sivertson S. A model for evaluating the history taking and physical examination skills of medical students. Med Teach. 1989;11:75–80.
47 Burch VC, Seggie JL, Gary NE. Formative assessment promotes learning in undergraduate clinical clerkships. S Afr Med J. 2006;96:430–433.
48 Haertel E. Getting the help we need. J Educ Meas. 2013;50:84–90.
49 Dweck CS. Motivational processes affecting learning. Am Psychol. 1986;41:1040–1048.
50 Lineberry M, Soo Park Y, Cook DA, Yudkowsky R. Making the case for mastery learning assessments: Key issues in validation and justification. Acad Med. 2015;90:1445–1450.
51 Lord SJ, Irwig L, Simes RJ. When is measuring sensitivity and specificity sufficient to evaluate a diagnostic test, and when do we need randomized trials? Ann Intern Med. 2006;144:850–855.
52 American Board of Medical Specialties. Standards for the ABMS program for maintenance of certification (MOC) for implementation in January 2015. http://www.abms.org/pdf/Standards%20for%20the%20ABMS%20Program%20for%20MOC%20FINAL.pdf. Accessed December 15, 2015.
53 Cronbach LJ. Five perspectives on validity argument In: Wainer H, Braun HI, eds. Test Validity. Hillsdale, NJ: Routledge; 1988:3–17.
54 Lane S. Validity evidence based on testing consequences. Psicothema. 2014;26:127–135.

Reference cited in Table 2 only

55 Huang GC, Newman LR, Schwartzstein RM, et al. Procedural competence in internal medicine residents: Validity of a central venous catheter insertion assessment instrument. Acad Med. 2009;84:1127–1134.

Appendix 1

A Framework for Organizing Sources of Consequences Evidence for Educational Assessments, With Examples[a]

Domain of impact	Specific consequence	Example (hypothetical) studies
Impact on examinee		
Topic-specific KSB	From score use: KSB resulting from additional training directed by assessment results (e.g., mastery learning, remediation)	Single-cohort study provides remediation to low-scoring students, and then tracks their subsequent performance over the next 12 months.
	From activity: KSB gained while taking the assessment (test-enhanced learning) KSB resulting from altered learning behaviors before the test (e.g., cramming)	Randomized trial of test vs. no test; outcome is KSB retained one week after test.
Non-topic-specific (noncognitive) behaviors	From score use: Improvements in general study or test-taking skills directed by assessment results (e.g., non-topic-specific remediation).	Students in lowest quartile on an end-of-year progress test are required to attend a six-hour study skills course; performance in the following academic year is tracked and compared.
	From activity: Learning behaviors (study patterns before or after test; self-regulation; study organization; additional training / remediation) Communication, teamwork Cheating	Randomized trial of midterm exam vs. no midterm; outcome is use of online learning resources as a marker of self-directed learning behaviors. A qualitative study finds that practicing nurses who prepare as a group for a newly required resuscitation skills assessment report greater team cohesiveness.
Motivation	From score use: Motivation to learn (intrinsic vs. extrinsic, goal- vs. performance-oriented, etc.)	Historical-control study: Investigators give students normative test score data (e.g., percentile rank within their cohort), making comparison with students from previous academic year who did not receive normative scores; outcome is how strongly performance-goal orientations are adopted after the test.
	From activity: Same as above	Pre–post study, comparing intrinsic interest in the topic one week before and immediately after the test (before scores are released).
Social and professional status	From score use: Status among peers Progress through program or career (remediation, promotion, certification) Completion/dropout rate Special recognition Employment Professional identity	Single-cohort study provides remediation to low-scoring students, then performs one-on-one interview at six months to explore how this affected their peer relationships and sense of well-being. Investigators seek to determine whether an assessment meant to be formative inadvertently influences clerkship directors' decisions as to which learners will earn "honors" grades. They conduct one-on-one interviews with clerkship directors about their grading decision process and then perform content analysis of transcripts.
	From activity: Completion/dropout rate Employment Professional identity Professional status	Survey study finds that a small percentage of physicians have let their board certification lapse solely because they didn't want to retake the high-stakes board exam.
Acculturation	From score use: Values (e.g., relative importance of tested topics)	Observational study finds that preceptor ratings on their clinical clerkships are the second most important predictor of specialty choice.
	From activity: Values Commitment and sense of belonging (e.g., to profession, institution, work unit)	Longitudinal study of physician trainees evaluates the association between high-stakes licensure and certification exams focusing on medical knowledge, completed during school and residency, and perceived importance of communication and interpersonal skills in their professional role. Qualitative analysis of focus group transcripts finds that physicians ascribe deep personal meaning to their completion of arduous high-stakes examinations, and that such completion is a catalyst for bonding within their profession and for differentiation between specialties.
Emotions and well-being	From score use: Self-concept (e.g., pride, self-actualization) Stress, anxiety Mood (e.g., depression)	Study finds that medical students in lowest quintile on USMLE Step 1 have a sevenfold higher rate of suicide attempts over the following year.
	From activity: Same as above	Time-series study tests how examinees' salivary cortisol response immediately prior to, during, and after a high-stakes board examination compares to that observed among subjects prior to, during, and after other high-stress endeavors (e.g., skydiving).

(*Appendix continues*)

Appendix 1
(Continued)

Domain of impact	Specific consequence	Example (hypothetical) studies
Resources (time, financial, other)	From score use: Time (i.e., time dedicated to test preparation or additional training) Finances (e.g., cost of assessment [registration] and associated expenses [travel])	Longitudinal study evaluates the time and money spent on board review courses for those who fail their first certification exam attempt.
	From activity: Same as above	Randomized trial of midterm exam vs. no midterm; outcomes include self-reported time spent studying in week leading up to exam.
Impact on educator		
Topic-specific KSB	From score use: Curriculum/content planning (e.g., teaching to the test) Instructional approach (e.g., improvements in response to low scores)	Three years after implementation of a locally developed test required for advancement to clinical training, investigators find that scores have improved by 23%. However, scores on other tests (including commercially developed tests) have not changed, suggesting that the improvement might be due to focused teaching/coaching or studying ("teaching to the test").
	From activity: Curriculum/content planning (e.g., teaching to the test)	A national licensing board initiates a substantial change in the exam structure. Researchers conduct a nation-wide survey to quantify the relative distribution of course content across all accredited schools just after this change is announced (prior to implementation), and repeat the survey two years after implementation.
Assessment skills	From score use: Accuracy of impressions about learners Ability to assess learners (topic-specific or general)	Rater training is offered to teachers who submit substandard clinical rotation ratings (scores above or below average, or scores with little between-student variability), and the effectiveness is evaluated.
	From activity: Same as above	See example in Table 2 (Impact on educators: Assessment and feedback skill).
Non-topic-specific (noncognitive) behaviors	From score use: Cheating (discouraging or condoning)	Following a statewide mandate requiring successful completion of annual progress testing for medical students, a survey finds that 32% of instructors turn a blind eye to students who secretly use notes during the written test.
	From activity: Teacher collaboration	Investigators use social network analysis to examine the degree to which continuing medical education instructors form and engage in networks to create curricula and share ideas. They conduct an existing-groups study comparing specialties in which MOC requirements have greatly intensified vs. those with relatively minor MOC changes.
Social and professional status	From score use: Peer and public perception of teacher Special recognition (e.g., if students perform very well or very poorly) Teaching assignments	Survey and interview study of basic science course faculty asks them to consider varying distributions of student assessment performance, and describe what recognition or reprimand they would anticipate from the dean. Investigators conduct a similar evaluation with leaders and administrators in the dean's office. Instructors whose students perform in the top 20% on their surgery clerkship are selected to staff a new teaching-only clinic; impact is evaluated by comparing performance of medical students who do or do not rotate through this clinic.
	From activity[b]	
Acculturation	From score use: Values (e.g., relative importance of tested topics) Commitment and sense of belonging (e.g., to profession, institution, work unit)	Survey of physicians who left academia finds that the fifth most common reason for leaving was the incongruence between espoused and demonstrated values in handling students in need of remediation.
	From activity: Same as above	Longitudinal survey investigates nursing instructors' perceived importance of communication skills before and after the introduction of a high-stakes communication skills exam.
Emotions and well-being	From score use: Self-concept (e.g., pride, self-actualization) Stress, anxiety Mood (e.g., depression)	Study finds that teachers whose students score below the national average show burnout twice as often as their peers.
	From activity: Same as above	Focus group study exploring sources of teacher stress finds that final exam creation and administration is a dominant life stressor.
Resources (time, financial, other)	From score use: Time (i.e., time dedicated to curriculum updates or additional training) Finances (e.g., financial incentives associated with students performing well)	The dean of a new medical school offers faculty a 10% salary bonus if average board scores are in the top quartile; intended and unintended impacts at three years are evaluated in a mixed-methods interview study.

(*Appendix continues*)

Appendix 1
(Continued)

Domain of impact	Specific consequence	Example (hypothetical) studies
Resources (time, financial, other), cont'd	From activity: Time Finances (e.g., financial incentives arising from test preparation or administration)	National survey finds that instructors spent, on average, 47 hours modifying their course to accommodate a change in the focus of a national certification exam. Survey study quantifies the income received by medical school teaching faculty for developing board questions, administering oral boards, or teaching board review courses.
Impact on other systems and people		
Targets of KSB application	From score use: Patient health Function of health care teams Function of health care systems	Observational study examines associations between board scores and the frequency of physician errors.
	From activity: Same as above	Because of increasingly stringent MOC requirements, 22% of physicians allow their board certification to lapse. Investigators use this information in conjunction with national safety metrics to evaluate the quality of care for certified vs. noncertified physicians.
School leaders and administrators	From score use: Curriculum/content planning (teaching to the test) Test security (prevent cheating) Leadership tenure in office	A national survey of all accredited medical schools examines the association between a school's average MCAT scores and the average duration of dean and associate dean tenure. See other examples above on curriculum planning and cheating.
	From activity: Same as above	See other examples above on curriculum planning and cheating.
Public and policy makers	From score use: Public perception of school Public perception of profession Public policies regarding training Public policies regarding uses of assessment scores	Observational study finds that after licensing exam results for each medical school are made publicly available, schools in the lowest tertile note a 27% drop in applicants over the next three years, and average MCAT scores drop by six percentile points.
	From activity: Same as above	Maintenance-of-certification organizations drop all assessments other than an every-10-year written exam. In response to public outcry, state governments assume responsibility for certifying physicians in 36 states over the next four years.
Others	From score use[b]	
	From activity: Untested students Test preparation programs	Following implementation of a new program of regular direct observation and feedback in the clinical year of a physical therapy training program, exam scores in the preclinical years drop by 7%. Focus groups with students and instructors suggest that faculty members are focusing more time on clinical students, and as a result preclinical students are neglected. Researchers investigate the economic impact of changing the format of the MCAT. They find that test preparation companies increased their operating budget by more than 30% over two years to accommodate new procedures and test content, that book publishers launched a new series of test preparation books, and that as a result students' out-of-pocket preparation costs increased by 43% in comparison with historical cohorts.
Impact of classifications[c]		
Labels for test and subtests	Interpretations/meaning inferred by users	Student focus groups suggested that exam results phrased as "Fail" led to reduced self-efficacy and self-image. Leaders used a quantitative survey to establish a baseline, then changed reporting to "Needs improvement" and repeated the survey one year later.
Topic-specific KSB score cut points (standard setting)	Pass/fail decisions Pass/fail rates	See example in Table 2 (Impact on defensibility: Establishment of passing standard). Instructors collect residents' scores on basic laparoscopic skill simulations just prior to beginning supervised surgical practice. To establish a pass/fail cut point, they also collect standardized assessments of surgical performance during residents' first three months of supervised practice, and use receiver operating characteristic curves to identify the cut point for basic skills that predicts acceptable supervised performance. Over the next three years, initial failure rates vary from 5% to 12%, which program directors feel is about right. Researchers seek to confirm the appropriateness of a new rigorously established cut point for a simulation-based assessment of LP skill, which is 12% higher than the current standard (i.e., more residents likely to fail). They randomly assign residents at 10 institutions to meet either the current or new cut point before performing LPs on patients, and study outcomes of volume of LPs performed by residents and rate of LP complications, each measured for the year prior and two years after.

(*Appendix continues*)

Appendix 1
(Continued)

Domain of impact	Specific consequence	Example (hypothetical) studies
Topic-specific KSB differential score functioning across subgroups	Bias in test scores (construct-irrelevant variance)	A differential test functioning study shows that women score lower on a novel test of clinical reasoning, even though their performance on other measures (including locally developed knowledge tests, clinical performance assessments, and national licensure exams) are similar. A follow-up exam using a commercial computer-based test of clinical reasoning confirms similar reasoning abilities, suggesting that the response format of the novel test is inadvertently biased against women.

Abbreviations: KSB indicates knowledge, skills, behaviors; USMLE, United States Medical Licensing Examination; MOC, maintenance of certification; MCAT, Medical College Admission Test; LP, lumbar puncture.

[a]Consequences evidence can be organized by whether it focuses on the impact on people or organizations, or the impact of classifications (first column). Evidence of impact can be further classified as arising directly from the use of scores or from the assessment activity itself (second column). All examples are hypothetical.

[b]The authors could not think of a plausible, meaningful example of evidence for this issue, but that does not preclude the possibility that others could identify such.

[c]The impact of classifications might be better viewed as preconsequences, with downstream impact on other true consequences noted above.