Validation of ELA and Mathematics Assessments: A General Approach[1]

Joan L. Herman & Kilchan Choi

Overview

Validity refers to the degree to which an assessment actually measures what it claims to measure and well serves intended purposes. From this perspective, assessments themselves are not valid or not; rather evidence of validity must be established in the context of specific interpretations and uses of test scores. A test may be well suited for one use and not for another. For example, a mathematics test appropriate for accountability or summative purposes may not well serve formative assessment purposes for teachers. Moreover, validity is a matter of degree; validation requires the *accumulation of evidence* to support the argument that scores derived from a given test yield accurate inferences to support intended interpretations and uses (AERA, APA, NCME, 1999; Kane, 2001). Finally, our validity definition requires consideration of both (1) what is measured—that an assessment measures what it is intended to measure—and (2) what interpretations and uses the assessment is intended to serve. For both summative and formative assessments, the definition implies that assessments must yield technically sound measures of student learning. Moreover, results should be useful and used for intended purposes and not carry serious unintended or negative consequences.

Modern theory suggests that validation be approached as an argument that needs to be substantiated. A validity argument lays out the claims that an assessment and its scores must satisfy to serve its proposed purpose(s) and/or use(s). Validation efforts then focus on the collection of evidence to document how well the assessment satisfies each claim. Below, we set out these claims as a set of criteria that educational assessments used for summative or accountability purposes should meet and present plans to systematically collect evidence to evaluate these claims throughout our test development and field-testing process.

> Validity is a matter of degree; validation requires the accumulation of evidence to support the argument that scores derived from a given test yield accurate inferences to support intended interpretations and uses.

Validity Criteria

The criteria combine the *Standards for Psychological and Educational Testing* (AERA, APA, & NCME, 1999) with the National Research Council's core dictum that assessments intended to both measure and benefit learning must first and foremost be themselves learning-based (Pellegrino, Chudowsky, & Glaser, 2001). This means that in addition to adhering to the *Standards'* core concepts of validity, accuracy and reliability, comparability, and fairness, assessments should fully reflect the intended learning domain and a model of how such learning develops. In the current national context, the Common Core State Standards define the prevailing domain specification and figure prominently in our validation approach. We draw on CRESST's ontology methodology—representations of major idea and principles that define a subject matter domain, the enduring concepts, knowledge, skills and attendant cognitive demands that constitute understanding of each, and the inter-relationships among them (see Baker et al., 2009; Iselli, 2010)—for both documenting the alignment of standards and assessment and, over time, validating and refining essential paths for learning.

National Center for Research on Evaluation, Standards, & Student Testing
UCLA | Graduate School of Education & Information Studies

The intent that both accountability and formative assessments will support the improvement of learning also brings with it the need for tests that are instructionally sensitive and have value for teaching and learning. Instructional sensitivity – that test scores reflect the quality of instruction, rather than home background and/or native ability—is a fundamental assumption underlying current accountability policy. If teaching improves, test scores will do likewise – otherwise, it is not reasonable to hold teachers and schools accountable for performance. Similarly, if scores are to be used to assess student progress – for example, the extent to which students are on track to college readiness—then scores must be comparable across time and vertically scaled.

Validity, in short, resides in evidence of the extent to which an assessment embodies the characteristics that support its intended purpose(s) and the extent to which the scores from an assessment yield meaningful inferences to support intended decisions and uses. These concerns suggest two basic kinds of criteria for judging test quality: one set based on the attributes or the characteristics of the test itself, and the other based on the validity of score interpretations. We offer an initial list in *Table 1* and *Table 2* below:

Table 1: Criteria Related to Assessment Attributes & Validity of Score Interpretation

Criteria Related to Assessment Attributes
Learning Based: • Aligned with significant learning goals, as reflected in Common Core State Standards and represented through CRESST ontologies, including expected content and cognitive demand • Comprehensive in representation of target constructs/domain intended for assessment, in both content and cognitive demands • Linked to expected trajectories of learning
Fairness • Accessible, enabling all students to show what they know • To the extent possible, free of knowledge and skills that are irrelevant to the target of the assessment, (e.g., construct irrelevant language demands) • Sensitive to a range of student abilities; appropriate for students at the range of developmental levels likely in assessed population • Provide accommodations where needed
Learning/Instructional Value • Incorporates transfer; new, authentic applications • Cognitively expansive (i.e., provides opportunities for students to organize and expand their knowledge, (e.g., through explanation, modeling, etc.) • Engaging • Consequences: models/supports rich curriculum learning opportunities; well communicates important learning goals
Utility • Timely • Useable/interpretable by teachers and/or students; provides actionable feedback for intended users • Instructionally useful, at the right grain size to guide subsequent, intended decision-making and action
Credibility • Educators • Students and parents • Public

Table 1: Criteria Related to Assessment Attributes & Validity of Score Interpretation (cont.)

Criteria Related to Validity of Score Interpretation
Technical Soundness • Score reliability at level of intended use(s) (e.g., if assessment is formative, scores provide reliable diagnostic information) • Reliability of scoring • Accuracy for intended decision-making purpose(s), including monitoring progress and determining attainment of standards
Fairness/lack of bias
Model/theory-based relationships with other measures
Generalizability, extent of transfer
Instructional sensitivity
Comparable, as necessary: across sites, across time, within and across years
Consequences • Supports productive changes in teachers pedagogical and assessment practices • Supports improvement in student learning • Supports equity in student outcomes.

Validation Plan

Assessment validity cannot be an afterthought but rather needs to be built in throughout a systematic test development process. That process starts with clear delineations of the purpose of the assessment, the domains to be assessed, characteristics of the examinee population and intended users. Within existing constraints and resources, test blueprints can then be developed to specify the number and types of items to be sampled for specified content and cognitive demand domains. Task design/templates then may be developed or selected to reflect the blueprint specification and used to generate items to populate it. Once initially developed or selected, test items, tasks and test forms go through multiple levels of review and trial to fine-tune and assure quality and to generate evidence of validity. Just as assessment has been described as a process of "reasoning from evidence (Mislevy et al. 2002; Pellegrino et al, 2001)," so too is assessment development an evidence-based process.

Mirroring the two general types of criteria describe above, our validation plans are designed to triangulate evidence from two basic frames: (1) expert and educator review and (2) empirical study. For example, subject matter, learning and cognition, and language experts can be convened at multiple stages of the test development process to examine the alignment between the assessment to the Common Core State Standards (as represented in content and cognitive demand ontologies), the assessment's incorporation of research-based design principles to directly enhance learning, and to identify/analyze construct irrelevant bias that may be introduced by linguistic complexity or cultural differences. Drawing on the teachers involved in the empirical validity studies described below, feedback on issues of utility, usability and feasibility can be solicited. A sensitivity review by multi-cultural experts also is planned. Review strategies figure prominently in the early states of test development; empirical studies are built into pilot, field tests, and continuing study of operational use and consequences. These plans are summarized in *Table 3*.

> "Just as assessment has been described as a process of "reasoning from evidence," so too is assessment development an evidence-based process."

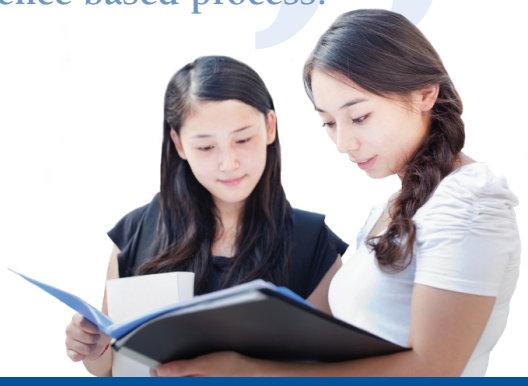

Table 3: Validation Plan

Test Development Step	Validity Criterion	Design/Data Source
1. Test Specification	**Learning-based:** • Alignment with standards/ontology • Comprehensiveness	**Expert review**
2. Task specifications (Design), including initial scoring rubrics	**Learning-based:** • Alignment with standards/ontology • Linked to expected trajectory	**Expert review**
	Fairness	**Pedagogical/educator review**
	Learning/Instructional value	**Educator review**
	Feasibility/credibility	
3. Task Development/Developmental Testing	**Learning-based:** • Alignment with standards/ontology • Linked to expected trajectory	
	Fairness	**Sensitivity/accessibility reviews**
	Learning/Instructional value	
	Feasibility/credibility	**Empirical trial: 10-50 students from target grade levels for each task— practicality, timing, clarity**
4. Field Test (test forms/sets of items and tasks)	**Learning-based:** • Alignment • Comprehensiveness • Relationships with other measures	**Expert review** **Empirical study: *n* = at least 20 classes and 400 students, representing a full range of student ability.** • Oversample depending on intended focal groups for DIF analysis • Each student responds to at least two tasks • 4-6 scorers • Supplement with "expert sample," *n* = 50-100. • Teacher data: OTL, utility, credibility, feasibility • Other concurrent student data: grades, standardized test, demographics
	Technical Soundness • Score reliability at level of intended use • Scorer reliability	
	Fairness/lack of bias	
	Generalizability	
	Instructional sensitivity (preliminary)	
	Model relationships • Concurrent validity • Expert-novice comparisons • Overall vs. subgroup models	
5. Operational Use	**Instructional Sensitivity**	**Empirical studies: *n* = 80 classrooms randomly assigned to instructed and non instructed groups; 3 tasks; data as above** **Other variations: through course/end course administration** **Teacher data: utility, credibility, feasibility** **Instructional practice data** **Longitudinal sample across three grades (*n* =300 students), data as above with external growth and success indicators.**
	Technical Soundness	
	Model Relationships as above	
	Teacher interpretation/use	
	Consequences/differential impact	
	Quality of implementation/teacher scoring	
	Validity of standard setting/vertical trajectory across grades	

References

AERA, APA, NCME. (1999). *Standards for educational and psychological testing*. Washington, DC: APABodilly, S., & Beckett, M. K. (2005). Making out-of-school time matter: Evidence for an action agenda. Santa Monica, CA: RAND.

Baker, E., Chung, G., & Herman, J. (2009). *Ontology-based educational design: Seeing is believing*. Los Angeles, CA: CRESST

Iseli, M. (2010). *Ontology system overview.* Los Angeles, CA: CRESST.

Kane, M. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, 38(4), 319–342.

Mislevy, R.J., Steinberg, L.S., Breyer, F.J., Almond, R.G., & Johnson, L. (2002). Making sense of data from complex assessment. *Applied Measurement in Education*, 15, 363-378

Pellegrino, J., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.