Journal of Writing Assessment Vol. 2, No. 1, pp. 65-70

An Annotated Bibliography of Writing Assessment
Reliability and Validity, Part 2

ELLEN SCHENDEL
Grand Valley State University

PEGGY O'NEILL
Loyola College

MICHAEL NEAL
Clemson University

BRIAN HUOT
Kent State University

In this, our third installment of the bibliography on assessment, we survey the second half of the literature on reliability and validity. The works we annotate focus primarily on the theoretical and technical definitions of reliability and validity—and in particular, on the relationship between the two concepts. We summarize psychometric scholarship that explains, defines, and theorizes reliability and validity in general and within the context of writing assessment. Later installments of the bibliography will focus on specific sorts of assessment practices and occasions, such as portfolios, placement assessments, and program assessment—all practices for which successful implementation depends on an understanding of reliability and validity.

As these annotations focus on technical and theoretical understandings of validity and reliability, and the two terms are often discussed in assessment scholarship together, we have not broken this installment of the bibliography into subsections. Furthermore, we have focused our attention on published scholarship of the field and have omitted unpublished sources such as ERIC documents and dissertations. We attempted to be thorough, but we hope readers will alert us to any omissions they note.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18*(2), 5-11.
Argues that validation must account for empirical and ethical information regarding decisions made based on test results. Validation includes both the existing evidence as well as speculation of future consequences associated with these decisions. Unified construct validity brings together content, criteria, and consequences of which none can stand alone

in the validation process because a full view of validity includes relationships between test interpretation, test use, evidence, and consequence.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: American Council on Education and Macmillan.
The definitive take on validity from the most quoted theorist. His definition of validity is exactly the same as that in the shorter version in *Educational Researcher* published the same year. His theories and general orientation are the same as well. In this version he considers validity from as many different perspectives as possible, recalling its history while at the same advancing his own arguments in traditional rival-hypothesis' discourse. The 79-page treatment (there is an 11-page bibliography—valuable in its own right) contains too many sections, headings and topics to even begin to list them all. Currently, this is the most complete and compelling argument for the concept of "Unified Test Validity."

Messick, S. (1995). The standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice, 14*(4), 5-8.
Presents validity as a unified concept that applies to all assessments and that "combines scientific inquiry with rational argument to justify (or nullify) score interpretation and use." Argues that given this understanding, it is useful for performance assessments to address different aspects of validity: content, substantive, structural, generalizability, external and consequential. Describes each aspect as well as relevant issues and sources of evidence.

Moss, P. A. (1994). Can there be validity without reliability? *Educational Researcher, 23*(2), 5-12.
Yes—assessments can be valid without being reliable. Argues for a hermeneutic approach to reliability, which recognizes the importance of context and "rational debate among the community of interpreters," as truth and interpretation are not fixed. Reliability should be one consideration among many in determining the usefulness and validity of an assessment. More important is whether the assessment contributes to teaching and learning and is ethical. By making reliability one concern among many—rather than the determining factor in an assessment's value—"the possibilities for designing assessment and accountability systems that reflect a full range of valued educational goals become greatly expanded."

Moss, P. A. (1998). Testing the test of a test: A response to the multiple inquiry in the validation of writing tests. *Assessing Writing, 5,* 111-122.
Critiques "Multiple Inquiry in the Validation of Writing Tests" for not following through with the placement illustration as an example of multiple inquiry validation. Argues that educational validity theory has been multimethodological for many years and that inviting stakeholders with opposing perspectives can provide a more productive reading of test data. Epistemological assumptions held by various stakeholders increases communication between stakeholders, improving the validation process.

Moss, P. A. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research 62*(3), 229-258.
An "integrative and critical review" of the of the literature on validity, specifically in the context of performance assessments. Section one summarizes the emerging consensus about validity among validity scholars that is not reflected in the *1985 Standards*, which includes centrality of construct validity, the consequences of assessments, and contexts

and generalizations; section two synthesizes the questions, criteria or evidence that has been used in traditional validity inquiry; and the final section reviews "concerns" about traditional validity inquiry that seem to privilege standardized assessments. Explains contributions of Messick's and Cronbach's work to the emerging consensus. Identifies the emerging views of validity and critiques of traditional psychometric approaches.

Moss, P. A. (1995). Themes and variations in validity theory. *Educational Measurement: Issues and Practice 14*(2), 5-13.
Identifies six significant questions about validity facing the measurement community in the revision of the *1985 Standards for Educational and Psychological Testing*. Articulates the tensions, contradictions and gaps between validity theory and practice. Argues that Cronbach and Messick's work in validity is considered "seminal" in any discussions of validity, and that the new *Standards* should reflect the "emerging consensus among validity theorists about the inadequacy of the construct—content-criterion framework for guiding validity research, about the centrality of construct validity to the evaluation of any assessment-based interpretation, and about the importance of expanding the concept of validity to include explicit consideration of the consequences of assessment use" (p. 12).

Moss, P. A., Beck, J.S., Ebbs, C., Matson, B., Muchmore, J., Steele, C.T., & Herter, R. (1992). Portfolios, accountability, and an interpretive approach to validity. *Educational Measurement: Issues and Practice, 11*(3), 12-21.
Argues for the use of classroom-generated performance assessments for accountability and large-scale assessments because standardized assessments can limit learning in significant ways, and classroom-based material can provide information that is inaccessible in other formats. Advocates for using interpretative research methods to analyze and report classroom-based assessments. Draws on F. Erickson's (1986) approach, which requires largely inductive, iterative methodology that includes a coding scheme and repeated testing of conclusions and assertions to account for all data. Results should be delivered in a research report, typically an interpretive narrative that contains particular descriptions, general descriptions, and an interpretive frame supported by evidence. Researcher/evaluator usually has a long-term involvement in the context and is a persistent observer. Data is triangulated with a clear evidentiary, audit trail that supports the conclusions. A detailed example using a Grade 8 Language Arts portfolio illustrates the approach.

Powers, D. E., Burstein, J. C., Chodorow, M. S., Fowles, M. E., & Kukich, K. (2002). Comparing the validity of automated and human scoring of essays. *Journal of Educational Computing Research, 26*(4), 407-425.
Reports on a study of the relationship between automated and human scoring of essays, and also the relationships between automated scores and non-test indicators of writing skills and scores generated by humans and non-test indicators of writing skills. Examines scores for the GRE Writing Assessment as generated by *e-rater*. Finds that human readers are reliable and that those scores are generally consistent with *e-rater* scores. Further finds that human scores correlate somewhat better to non-test indicators than do *e-rater* scores, but that there is potential for machine "scores as (valid) indicators of prospective graduate students' writing skills, especially when they can be combined with scores provided by at least one human reader." Points out that "although unlikely perhaps, it is possible that human and automated scores could both be based, at least in part, on features that are not entirely relevant to good writing" and that the validity of machine-generated scores needs to be measured against more than the scores generated by human readers.

Scharton, M. The politics of validity. In E. M. White, W. D. Lutz, & S. Kamusikiri (Eds.), *Assessment of writing: Politics, policies, practices* (pp. 52-75). New York: Modern Language Association.
In describing the political nature of validity, defines terminology such as *construct validity, content validity, concurrent validity, predictive validity,* and *interrater reliability.* Argues that there is a tendency in educational assessment to argue for validity based on its theoretical savvy or professional currency. Warns that the use of validity as a tool for ideological or social change within education will undermine the purpose(s) of the assessment at hand.

Shepard, L.A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice, 16*(2), 5-24.
Test validation should include measurement of test effects and social justice. Agrees with Messick's conception of a unified validity, although takes issue with his matrix showing the relationships between test interpretation, test use, evidential basis, and consequential basis, arguing that it visually separates construct validity from social consequences. Coaching on tests demonstrates not only results being manipulated and thus producing invalid scores but also a flaw in test design that allows for such gains to be made, which she expects test makers to deal with more adequately. On the other hand, she distinguishes test misuse from side effects for which test makers should not be blamed. When tests are used for a new purpose, a fresh validity evaluation is required.

Shepard, L. A. (1993). Evaluating test validity. *Review of Educational Research in Education, 19,* 405-450.
Provides a historical overview on the development of current understandings of validity. Explains the traditional tripartite view (content, criterion, and construct) of validity and its rejection in favor of a unified approach under construct validity. Presents contributions of Cronbach and Messick, emphasizing Messick's unified theory and how it challenges these traditional views, specifically in terms of consequences. Argues for a reformulating of Messick through a central question—"What does the testing practice claim to do?"—which would be used to identify priorities and further questions for investigation. Describes the gap between validity theory and practice and offers four examples to illustrate approaches to validity inquiry that fit within the proposed framework. Concludes with the implications of this approach to validity for validity researchers and test evaluators as well as the revision of the *1985 Standards.*

Smith, W. L. (Ed.). (1998). Validation and writing assessment. [Special issue]. *Assessing Writing, 5*(1).
This special issue of the journal includes an introduction by Smith and three essays on different aspects of validation, each accompanied by a shorter response essay. The first essay, "Writing Assessment: Raters' Elaboration of the Rating Task," by Mary L. DeRemer, uses think-aloud protocols to examine how raters constructed specific tasks as they evaluated student portfolios with a analytic scoring rubric. DeRemer concludes that the raters "engage in extensive problem-solving activity." Raters also constructed different task elaborations that affect the meaning of the scores assigned, which has implications for validity. Harry Torrance, who responded to this essay, acknowledges the potential contribution the article makes to the under-researched topic of raters as they are rating, but critiques the essay for methodological flaws.

The second article, "A Question of Choice: The Implications of Assessing Expressive Writing in Multiple Genres," by Gail L. Goldberg, Barbara S. Roswell, and Hillary

Michaels, examines one task of a statewide mandated writing test using analysis of scoring data, questionnaires, student interviews, and textual analysis of 300 student texts. The authors argue "strongly for the validity of this choice task as a measure of expressive writing" and contend that this genre "increases writers' engagement and enhances the fairness of the assessment by giving all students the best opportunity to demonstrate proficiency of this outcome." In their response, Roger D. Cherry and Stephen P. Witte use "A Question of Choice" as a springboard to discuss other issues associated with writing assessments, such as the significance of context in writing and definitions of writing ability. They conclude that "to enhance the validity of direct assessments of writing, writing assessment practitioners must develop an understanding of writing ability that better accommodates the ways in which written documents function" in naturally occurring contexts of text production and use.

Richard H. Haswell's, "Multiple Inquiry in the Validation of Writing Assessments," uses the case study of a writing placement exam as a "test case" for a multimethod validation process. Haswell argues that multiple lines of inquiry are needed "to be of use to a variety of stakeholders, to be sensitive to the presence of conflicting perspectives, to seek convergent findings . . . and to probe a social context that is complex, fluid and provisional." In "Testing the Test of the Test," Pamela Moss agrees with Haswell's basic premise. However, she critiques Haswell—and college writing assessment in general—for being "seriously isolated from the larger educational assessment community." She argues that there is already substantial research and theory in the educational measurement literature to support Haswell's argument, and his characterization of psychometrics isn't accurate. She concludes by encouraging Haswell and others to move beyond "conventional practices" and engage in critical reflection of their own theories and practices.

The final essay in the issue is "Validation of a Scheme for Assessing Argumentative Writing of Middle School Students," by Stuart S. Yeh, which reports on two studies that examined "factors influencing ratings of argumentative essays" in order to develop an analytic scheme for assessing essays. Yeh's work is grounded in Toulmin's theory of argument. Robert J. Bracewell responds positively to Yeh's methods and findings but argues that his scheme would not be very useful in looking at more complex argument structures (a fact that Yeh acknowledges). Bracewell offers a phrase structure tree of Toulmin's model, developed by J. G. Crammond, as an alternative analytical tool that allows for more nuanced analysis of texts.

Wiggins, G. (1993). The constant danger of sacrificing validity to reliability: Making writing assessment serve writers. *Assessing Writing, 1*, 129-139.
Writing assessments should reflect pedagogical values associated with current writing theory and practice as well as improve student performance. Writing assessments, even those that are performance based, inflate the value of reliability and efficiency over validity and thoughtful insight. Educators should be concerned about standardized writing prompts and rubrics that focus attention primarily upon the product and surface level characteristics of writing. Assessments instead should value inquiry, discovery, process, and meaning—characteristics that are consistent with the best approaches to the teaching of writing.

Williamson, M.M. (1994). The worship of efficiency: Untangling theoretical and practical considerations in writing assessment. *Assessing Writing, 1*, 147-174.
Efficiency, one of the dominant themes in American education since the end of the 19th century, is evidenced by the amount of money invested in schools, the work produced by teachers and administrators, and students' preparedness and productivity in the workforce. As the value of individual student merit rose in the 19th century, so did the use of

psychometric assessment tools, largely because they were understood to be fair and unbiased. Efficient technological innovations such as the multiple choice test and computerized scoring allowed for the development of large-scale, standardized tests that avoided writing, which was seen as too subjective and labor intensive to be viable. Suggests envisioning the teaching of writing as a "craft workshop" that values writing teachers' judgments and de-emphasizes assessments designed by those not involved in the teaching of writing on a regular basis.