Journal of Writing Assessment Vol. 1, No. 2, pp. 153-156

An Annotated Bibliography of Writing Assessment
Reliability and Validity, Part 1

ELLEN SCHENDEL
Grand Valley State University

PEGGY O'NEILL
Loyola College

MICHAEL NEAL
Clemson University

BRIAN HUOT
University of Louisville

In this, our second installment of the bibliography on assessment, we survey the literature on reliability and validity, the first of a two-part series that will continue in the next issue of *JWA*. 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 Furthermore, we have focused our attention on published scholarship of the field and have omitted unpublished sources such as ERIC documents and dissertations.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.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
Chapter 1 focuses on validity, and the 1999 standards give an updated definition, given the changing scholarship in the field. The chapter begins with this definition of validity: "Validity refers to the degree to which evidence and theory support the interpretations of

tests and scores entailed by proposed uses of tests." Rest of chapter explains the importance of validity and the technical aspects of establishing validity. Particularly helpful is the list of "Standards" for validity, which direct test administrators to examine particular aspects of assessment instruments and their usage.

Bachman, L.F. (2002). Alternative interpretations of alternative assessments: Some validity issues in educational performance assessments. *Educational Measurement: Issues and Practices, 21*(3), 5-19.
Shows how establishing validity of performance-based assessments is complicated, due to the complex nature of such assessment tasks, the fact that such assessments are meant to explore more complex issues, and the closer and more invested relationship between test administrator and test-taker. The complexity of the validity issue increases in assessments of language, as language is both the means for the assessment to occur and the thing being assessed. Argues two main points: "First, in both language testing and educational assessment, we must consider the roles of both language and content knowledge in the ways we define the constructs we want to measure, in the way we design assessment tasks, and in the kinds of inferences we can make on the basis of our assessments. Second, our approach to the design and development of language assessments and educational performance assessments must be both construct-based and task-based." Gives questions to consider when investigating the validity of alternative and performance-based assessments. Shows how, contrary to what the bulk of the literature on validity might say, "our approach [to language assessment and performance] must be both construct-based and task-based."

Berlak, H. (1992). Toward the development of a new science of educational testing and assessment. In H. Berlak, F. M. Newmann, E. Adams, D. A. Archbald, T. Burgess, J. Raven, & T. A. Romberg (Eds.), *Toward a science of educational testing and assessment* (pp. 181-234). Albany: State University of New York Press.
Locates current and traditional theories and practices for assessment in logical positivism and the search for objectivity and science. Discusses testing as a discourse that not only promotes a certain understanding of school and achievement, but that also produces its own discourses about school that locate power and control in the assessments and the scores. Outlines alternative assessment practices that recognize diversity in students, teachers, and achievement and locate power and control within individual communities.

Charney, D. (1984). The validity of using holistic scoring to evaluate writing: A critical overview. *Research in the Teaching of English, 18*(1), 65-81.
Positions analysis of the validity of holistic scoring within the debate between the use of quantitative, indirect assessments of writing, which tend to be reliable but perhaps not valid, and the use of qualitative, direct assessments of writing, which are said to suffer from validity. Argues against the notion that holistic scoring is the method that definitively settles this debate, as it is reliable (consistent) and valid (because it is a direct assessment of writing). Specifically, points out that holistic scoring may be less valid because of its reliability, as agreement may be arrived at by overexamining surface features of writing and not attending to more complex issues of writing and because the criteria used in holistic scoring is controversial, inconsistent within the field. Ends by arguing for a field-wide discussion of what constitutes "good writing" in order for holistic scoring criteria to be better fleshed out, therefore making writing assessments that rely on holistic scoring more valid.

Cherry, R., & Meyer, P. (1993). Reliability issues in holistic assessment. In M. M. Williamson & B. A. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (pp. 109-141). Cresskill, NJ: Hampton Press.
Gives a thorough overview of various technical issues related to reliability—such as instrument reliability versus interrater reliability and methods for estimating and reporting instrument and interrater reliability coefficients—in an effort to educate the profession on the matter of reliability and its relationship to validity. In defining terminology, explaining how aspects of reliability function within assessment, and giving specific information about how to check for and report on reliability, the chapter describes the ways in which the literature on holistic scoring often misinterprets or oversimplifies reliability, making it seem "fixed" or "monolithic" when it is in fact complex. Cherry and Meyer argue that writing assessment literature should be more detailed in explaining how reliability is arrived at, in order that "results can be compared across studies . . . [and] move toward informed, intelligent answers to important questions that remain about the direct measurement of writing skills."

Cherryholmes, C. H. (1988). *Power and criticism: Poststructural investigations in education*. New York: Teachers College Press.
Chapter 5 is titled "Construct Validity and the Discourses of Research." It discusses the historical evolution of construct validity from a postmodern perspective in which "Construct validity focuses on the juncture of words and things, concepts and objects, theory and practice, where social theory and research and theoretical constructs and research operations converge and diverge." Starting with the early work of Cronbach, it examines construct validity through phenomenological, critical, interpretive analytical and deconstructive lenses. Emphasizes that validity is a discourse of power and persuasion, linking it to social research and interrogation.

Crocker, L. (Ed.). (1997). [Special Issue]. *Educational Measurement: Issues and Practices, 16*(2).
This special issue takes on the "great validity debate," which Crocker says "has been brewing in psychometric circles for nearly a decade" and is intertwined with the revision of the *APA Standards*. The debate is between two views of validation: one sees it strictly as an empirical, scientific enterprise whereas the other includes the sociopolitical process (specifically in terms of the consequences of an assessment) along with empirical evidence. Four experts contribute to the debate in this issue: Lorrie Shepard, James Popham, Robert Linn, and William Mehrens. Shepard, grounding her position in the work of Messick and Cronbach, argues that consequential validity coincides with "long-standing principles of validity theory." Popham provides the counterpoint to her position, arguing that although test use consequences are important, they should not be included in validation. Linn closely aligns himself with Shepard's position and critiques Popham's argument. Mehren's response critiques Shepard's position, arguing for the measurement community to narrow the use of the term validity, not to enlarge it.

Cronbach, L. J. (1989). Construct validation after thirty years. In R. L. Linn (Ed.), *Intelligence measurement, theory and public policy: Proceedings of a symposium in honor of L. G. Humphreys* (pp. 147-171). Urbana and Chicago: University of Illinois Press.
Reviews the history of construct validity as it emanates from the germinal work done by Cronbach and Meehl in the 1950s. Furnishes an intellectual history of the thinking that led to the current emphasis on construct validity. Provides a look at the ideas that eventually supported construct validity as the most important criterion for making decisions

about tests. As well, each set of ideas is also linked to various examples of test development and use. A powerful retrospective from one of the minds behind validity theory as we now know it.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 3-17). Hillsdale, NJ: Erlbaum.
This is probably Cronbach's most definitive statement on validity. Along with Messick's work, Cronbach's definition of validity in this essay is the most widely cited in the literature. Validity is discussed in terms of its functional, political, economic, and explanatory perspectives. For Cronbach, validity is about making an argument for each use of a test. As a rhetorical construct, validity is contextual and partial, more or less compelling for different audiences and situations. Like Messick, Cronbach emphasizes the decisions made based upon a measure and the implications and consequences of any measure on the people and environment affected by the act of testing.

Greenberg, K. (1992). Validity and reliability issues in the direct assessment of writing. *WPA: Writing Program Administration, 16*, 7-22.
Rehearses many of the issues involving reliability and validity that fueled arguments about the use of direct and indirect measures for assessing writing. The brief history of writing assessment is detailed and connected to the then-current notions of validity and reliability. Provides an intellectual history of writing assessment that demonstrates how varying notions of writing, its assessment, and the concepts of validity and reliability combine to propel writing assessment away from multiple-choice tests of usage, grammar and mechanics.

Haswell, R.H. (1998). Multiple inquiry in the validation of writing tests. *Assessing Writing, 5*, 89-109.
Argues that because underlying assumptions about language testing have changed in the past two to three decades—from "universal aptitude" to public good—so must the way educators measure test validity. Measuring the social consequences of language tests requires a mixture of methods to collect and process information. Outlines four underlying principles for multiple inquiry: (a) educators use tests for different purposes, and each requires separate validation; (b) multiple methods include a wider array of stakeholder voices to combat privileging one particular perspective; (c) cross-checking or triangulation allows for a critical examination of biases and/or weaknesses, and (d) the probing nature of multiple inquiry assumes that no single validation study is ever complete. Details a multimethod approach to validation on a large-scale writing placement assessment, recommending that multiple stakeholders provide input on the process and interpretation of the outcomes.

Huot, B. A. (1990). Reliability, validity and holistic scoring: What we know and what we need to know. *College Composition and Communication, 41*, 201-213.
Although it foregrounds holistic scoring, the article summarizes understandings and uses of reliability and validity and, in particular, their relationship to one another. Gives overviews of the literature on reliability and validity, arguing that "the most important side effect of the constant stress on reliability is that it has caused the professional to assume, confuse, and otherwise neglect the validity of holistic scoring."

Journal of Writing Assessment

Index Volume One 2003

Articles

Bibliography

Reviews