Most of our knowledge about what humans can do, or tell us about their attitudes toward what they can do, may be expressed in two related concepts:

\[ \theta = T - E \]  
\[ C_\theta = D + W + Q \]

In the first, let \( \theta \) stand for a construct, a theory-based concept of behavior that varies according to context. Let \( T \) stand for the score, the value we give that construct as we observe it in a performance. Let \( E \) be the unbiased error of that measurement as we observe the construct. Hence, performance regarding the construct of interest is equal to the true score minus error. In the second, let \( C_\theta \) stand for the claim, based on the score, that we make about the construct identified in the first concept. Let \( D \) stand for the data, the information used...
to make the claim about the construct; let W stand for the warrant, the chain of explicit or implicit reasoning, that allows a relationship to be established between claim and supporting data; and let Q stand for the qualification, the limits in the degree of force expressed to validate the claim. The qualified claim we make about the construct is therefore supported by data explicitly or implicitly related to the claim.

Biographically, the origin of Concept 1 is found in the works of Karl Pearson, Charles Spearman, Edward L. Thorndike, Philip J. Rulon, Lee J. Cronbach, Anne Anastasi, Frederick M. Lord, Samuel J. Messick, and Michael T. Kane. Historically, the pursuit of validity is a product of the industrial revolution, a period in which the drives for technological innovation and economic efficiency were realized through creation of systems. Concept 2 finds its biographical origin in Stephen Toulmin. Historically a Cold War concept designed to acknowledge contingency, arguments of this sort—neither logically syllogistic (think Aristotle) nor remorselessly inductive (now Descartes)—signal that the old verities involving certainty vanished with the Habsburg Empire. A post-industrial creation, the particular genre of argument reified in the second concept has carried the first to our doorstep. Taken together, the two concepts may be viewed as metaphors for the ways that Western culture pursues certainty.

In remarkably enduring fashion, both the concept of validity and the arguments used to construct its presence have become ubiquitous creations. Whether managing receding capital or burgeoning information, whether used to justify a claim of gender discrimination based on salary, advance a clinical trial of a hypertension medication, or exempt a student from a writing class, the concepts adapt equally well—so well that we may wonder if they are not merely artifacts of the system but, rather, the System itself.

As they impact the field of education, the details are best found in the newly revised Standards for Educational and Psychological Testing. In its most recent form, the sixty-year-old evolving system of educational measurement is relevant for writing studies because it informs our empirical research paradigms including, but not limited to, the assessment of writing ability. Analysis of the Standards, conversely, is equally important as a way to understand the extent to which writing studies has influenced the system itself. Discussion of writing skills, holistic and analytic scoring, and automated writing evaluation are used throughout the sixth edition to demonstrate the strengths and weaknesses of the foundations, operations, and applications of the Standards. Because of fascination with the construct of writing as a remarkable human behavior of
enormous complexity, interchange has increased between the fields of educational measurement and writing studies. To understand the significance of the 2014 edition of the *Standards*, we must begin with what Cheryl Glenn and Jessica Enoch elegantly term the historiographic trajectory of concepts that largely determine the way we think now.

**Origin: From Technical Recommendations to Measurement System**

In 1906 a committee of the American Psychological Association (APA) chaired by James Rowland Angell of the University of Chicago was appointed to standardize measurement techniques associated with experimental research. In 1952 another APA committee, this time with the American Educational Research Association (AERA) and the National Council of Measurement in Education (NCME, formerly the National Council on Measurements Used in Education), was appointed to standardize the reporting of information gathered from just those techniques. The shift in attitudes from uniformity of measurement in 1906 (a period in which the fabric of American life was systemized to provide solutions) to audience awareness in 1952 (when those social solutions had come to be viewed as problems) was accompanied by a growth in college enrollment. Between 1899 and 1900, total postsecondary enrollment was 237,592 at 977 institutions; between 1949 and 1950, total enrollment rose to 3,639,847 at 2,004 institutions (Snyder, table 23). Developed first by the College Board and later by the American College Testing Program, tests were being deployed to sort those who would, and would not, be welcomed into a glossy new future, as shimmering as the answer booklets used to determine admissions. The fields of psychology and educational measurement began to have their fingerprints all over each other. Accompanying such interaction and growth, transparency to stakeholders became increasingly important.

“The essential principle that sets the tone for this document,” the 1952 preliminary proposal for technical recommendations stated, “is that a test manual should carry sufficient information that any qualified user can make sound judgments regarding the usefulness and interpretation of the test” (462). Chairing a committee from his academic home at the University of Illinois, Lee J. Cronbach and his colleagues had drafted a comprehensive document articulating the responsibilities of those who create and use tests. With a sense of public accountability extending far beyond the clubby atmosphere of 1906, the 1952 technical recommendations aimed to cover not just tests
but all evaluations used in assessing psychological characteristics. By 1954, the
document prepared by Cronbach was revised by a committee chaired by
Jacob S. Orleans of the City College of New York, printed separately, and sold
for $1.00 for wider distribution.

There were subtle yet important differences between the two versions of
the recommendations. While the 1952 document made no mention of construct
validity, the 1954 version, as Table 1 shows, introduced this concept along with
content validity (how well the measure at hand sampled the targeted subject
matter), predictive validity (how well predictions made from the measure are
confirmed by subsequent evidence), and concurrent validity (how well the
scores from the measure correspond to some other criteria known to evaluate
the subject matter). Focusing on procedure more than definitions, the recom-
mandations established that “construct validity is evaluated by investigating
what psychological qualities a test measures, i.e., by demonstrating that certain
explanatory constructs account to some degree for performance on the test.
To examine construct validity requires both logical and empirical attack. Es-
sentially, in studies of construct validity we are validating the theory underlying
the test” (14). Positioned to be the heart of the matter, construct validity was to
become the first among equals in the search for meaningful score interpretation.

The 1952 and 1954 documents also shared a statement of qualification
from the 1906 committee. Reservations about the bureaucratic nature of unifor-
mity served as a reminder that “it is necessary to make sure that standardizing
efforts do not stifle growth” (1952, 451; 1954, 1). Issuing specifications, the au-
thors warned, could discourage the development of new types of measurement.
While advancements had been made in measurement, the statement remained
as a warning that there were dangers in homogeneity. The way forward was
to be found in refusing to pit standardization against innovation: “It may be
noted that the wide variety of present tests is in itself proof that appropriate
standardization need not interfere with innovation” (1952, 462; 1954, 1). Such
dualisms, as history would demonstrate, were not so easy to resolve.

In 1955 another version of the recommendations was issued. In this ver-
sion APA had far less influence, and recommendations applicable to personality
inventories were withdrawn. While no warning label appeared, however, new
emphasis was placed on the audience—classroom teachers, counselors, supervi-
sors, administrators, and psychometricians—for a manual that accompanies
a test. In place of the standardizing committee (the maker) attention was now
paid to the individual or group (the user). It was a substantial shift.
Table 1. Evolution of Standards for Educational and Psychological Testing

<table>
<thead>
<tr>
<th>Title</th>
<th>Reference and Date</th>
<th>Sponsors and Key Authors</th>
<th>Definition of Validity</th>
<th>Innovation</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Technical Recommendations for Psychological Tests and Diagnostic Techniques: Preliminary Proposal”</td>
<td>American Psychologist 7.8 (1952): 461–475.</td>
<td>APA Committee on Test Standards, in consultation with AERA, NCME; Lee J. Cronbach (chair)</td>
<td>“Validity is not an absolute characteristic of a test. There are several types of validity depending on the type of inference for which the test is to be used. In the following standards four categories of validities have been distinguished; namely, predictive validity, status validity, content validity, and congruent validity” (467).</td>
<td>· Application to tests as well as published assessments for psychological diagnosis and evaluation&lt;br&gt;· Requirement that test manuals must contain information sufficient to enable a qualified user to make judgments regarding interpretation of a test&lt;br&gt;· Acknowledgment that standardization may stifle growth&lt;br&gt;· Introduction of three levels of recommendations (essential, very desirable, desirable)</td>
</tr>
<tr>
<td>Technical Recommendations for Psychological Tests and Diagnostic Techniques</td>
<td>Psychological Bulletin 51.2 (1954): 1–38; published separately in March 1954 as a supplement to Psychological Bulletin</td>
<td>Joint committee of AERA, APA, and NCME; Jacob S. Orleans (chair)</td>
<td>“Validity information indicates to the test user the degree to which the test is capable of achieving certain aims” (13).</td>
<td>· Emphasis on alignment of test with its stated purpose&lt;br&gt;· Introduction of construct validity</td>
</tr>
<tr>
<td>Technical Recommendations for Achievement Tests (1st edition)</td>
<td>National Education Association, January 1955</td>
<td>Committee on Test Standards, AERA; Committee of Test Standards, NCME; Jacob S. Orleans (chair, AERA); Robert E. Ebel (chair, NCME)</td>
<td>“Validity information indicates the degree to which the test is capable of accomplishing certain aims” (15).</td>
<td>· Limitation to achievement tests&lt;br&gt;· Emphasis on audience</td>
</tr>
</tbody>
</table>
Table 1. Continued

<table>
<thead>
<tr>
<th>Title</th>
<th>Reference and Date</th>
<th>Sponsors and Key Authors</th>
<th>Definition of Validity</th>
<th>Innovation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standards for Educational Tests and Manuals (2nd edition)</td>
<td>APA, 1966</td>
<td>Joint committee, AERA, APA, NCME; John W. French and William B. Michael (cochairs)</td>
<td>&quot;Validity information indicates the degree to which the test is capable of achieving certain aims&quot; (15).</td>
<td>· Restoration to broad scope of measurement · Introduction of trinitarian model validity: criterion, content, and construct</td>
</tr>
<tr>
<td>Standards for Educational and Psychological Tests (3rd edition)</td>
<td>APA, 1974</td>
<td>Joint Committee, AERA, APA, NCME; Frederick B. Davis (chair); Robert M. Guion</td>
<td>&quot;Questions of validity are questions that may properly be inferred from a test score; validity refers to the appropriateness of inferences from test scores or other forms of assessment. The many types of validity questions can, for convenience, be reduced to two: (a) What can be inferred about what is being measured by the test? (b) What can be inferred about other behavior?&quot; (25).</td>
<td>· Acknowledgment of social impact of score use · Introduction of investigation of bias through regression analysis · Definition of validity includes inference · Emphasis on interpretative meaning of test scores themselves · Attention to generalization and extrapolation inferences</td>
</tr>
<tr>
<td>Standards for Educational and Psychological Testing (4th edition)</td>
<td>1985</td>
<td>Joint Committee, AERA, APA, NCME; Robert L. Linn and Melvin R. Novick (cochairs)</td>
<td>&quot;Validity is the most important consideration in test evaluation. The concept refers to the appropriateness, meaningfulness, and usefulness of the specific inferences made about test scores&quot; (9).</td>
<td>· Acknowledgment that separation of scientific and social concerns is not possible · Introduction of two levels of recommendations: primary or secondary · Emphasis on particular applications for linguistic minorities and handicapping conditions</td>
</tr>
</tbody>
</table>

continued on next page
In 1966, the APA reentered as a full partner in the new edition under the direction of John W. French of the Educational Testing Service and Robert E. Ebel of Michigan State University. The emphasis on both educational and psychological measurement from the 1952 and 1954 editions was restored. While *Standards for Educational Tests and Manuals* adheres to essentially the same definition of validity that had prevailed earlier, its four categories that had been used in one form or another since 1906 were collapsed into what Robert M. Guion would later term the trinitarian model of criterion, content, and construct validity. As Robert L. Brennan has pointed out in his analysis of the 1966 *Standards*, validity was now linked to reasoning through inference.
Users do not merely apply scores; rather, they “infer the degree to which the individual possesses some hypothetical trait or quality (construct) presumed to be reflected in test performance” (12). Where once there was a test to be taken or left as a self-contained entity, now there were inferences to be made from the test scores. Because blanket statements were crude, to state that a test was invalid was no longer permissible. The last vestiges of certainty, trailing their glory through the Eisenhower administration, had been replaced by the complexities of Lyndon Johnson’s Great Society. From here on, it would be increasingly hard to claim the truth about anything.

This cultural shift is evident in the 1974 edition of the Standards. Under the direction of Frank B. Davis of the University of Pennsylvania, with contributions from Guion, this edition was distinctly, though awkwardly, attuned to social context. Civil rights legislation of 1964 and 1965 had perhaps been too current to impact the 1966 edition, but the concept of fairness is apparent in the 1974 edition. Accompanied by a palpable sense of defensiveness, equity of opportunity was given extended treatment: “It is sometimes suggested in response to perceptions of test abuse and unfair uses of tests that a moratorium on testing be observed until better and more appropriate instruments are developed and more equitable procedures can be instituted,” a lengthy note began. “The suggestion of such an extreme measure,” the note continued, “may be indicative of the growing sense of frustration and indignation felt particularly by some minority group members who sense that testing has had disproportionately negative impact on their opportunities for equal access to success” (2). The note then turned to the futility of even the suggestion that tests themselves might be unfair; it is the use of tests, the authors claim, not their existence, that is the problem.

The clumsiness of that logic—it’s about test use, not tests themselves—reveals a system uncertain of just how to account for itself. The 1961–1962 public attacks of Banesh Hoffman on testing in American schools, first in Harper’s and The American Scholar and then, more comprehensively, in The Tyranny of Testing, had led to a national consciousness increasingly critical of testing. By 1976, Ralph Nader was to declare to the New York Times, “If I have anything to do with it, it will be the end of tests” (qtd. in Fiske 404). Yet the more influential protest, internal to the APA, was the report of a committee chaired by Brent Baxter on testing of minority groups. Basing the report on the process of validation, Baxter and his colleagues questioned the steps used in the validation process. “What strategies may be useful,” they asked, “in learning whether or not particular items of information may differ in meaning for
particular individuals, especially those belonging to different subgroups such as disadvantaged and advantaged Negroes and whites” (646). Baxter’s questions struck at the heart of validity itself, especially in terms of its categories. Analysis of concurrent validity (how well the scores from an employment test, for example, correspond to job performance) may underestimate the true degree of relationship between test scores and job behavior if candidates for the job were initially rejected because of discrimination. The meaning of test scores for disadvantaged groups clearly deserved far greater attention.

Edginess about the Baxter report fuelled the lengthy note; defensiveness caused the awkward displacement of responsibility. On a technical level, however, the committee’s awareness of social context led to an important innovation regarding fairness. The 1974 edition of the Standards was the first to argue that regression analysis should be used to identify tests that “might be considered unfair to blacks (if they have a lower regression line) since their performance might be systematically under predicted” (44). By the following year, T. Anne Cleary and her colleagues would establish empirical comparison of regression as the chief method of identifying the potentially discriminatory impact of tests on minority groups. If the regression lines for test scores differed significantly between two groups, then differential prediction might be said to exist—an indication of test bias. The authors of the Standards had acknowledged, in a powerfully empirical way, the need to examine the meaning of scores across different groups of examinees. As a response to social protest, it was a landmark moment.

Two other editions were to follow—one in 1985, the other in 1999. The former was chaired by Robert L. Linn of the University of Illinois and Melvin R. Novick of the University of Iowa. Admitting the increasing pressure to acknowledge that social conditions influenced measurement, the introduction is straightforward in recognizing that “complete separation of scientific and social concerns is not possible” (1). The authors of a document designed to “provide a frame of reference to assure that relevant issues are addressed” were now more certain of their position. The three levels of recommendations of relative importance introduced in the first edition are now reduced to two (primary or secondary), and validity—“the most important consideration in test evaluation” (9)—remains primus inter pares. The organization of the volume into four parts was also new, with special attention given to “Testing Linguistic Minorities” (73–75) and “Testing People Who Have Handicapping Conditions” (77–80). The 1999 edition, cochaired by Eva Baker of the University of California, Los Angeles, and Paul Sackett of the University of Minnesota,
continued the emphasis on application, including a brief section on fairness in testing and test use (71–84). More widely peer-reviewed than previous editions, the 1999 edition received nearly eight thousand pages of comments from a wide variety of organizations, credentialing boards, government and federal agencies, test publishers, and academic institutions. As part of that review, the levels of recommendation were wisely abandoned altogether.

Significantly, the fifth edition deepened the definition of validity to include use of “evidence and theory” to support score interpretation and application (9). Unfortunately, the authors did very little with the provocative link between validity and theory. To examine that relationship, we must turn to the seminal chapter on validity by Samuel Messick in the third edition of Educational Measurement, published a decade before the 1999 Standards.

While reading Messick has been likened to carving a Thanksgiving armadillo (Markus), his prose becomes strikingly clear in dealing with the consequential basis of interpretation (60–63). His analysis of the value of theory, he plainly states, is “to raise consciousness about the pervasive consequences of value-laden terms (which in any event cannot be avoided in either social action or social science) and about the need to take both substantive aspects and value aspects of score meaning into account in test validation” (63). Test names, construct labels, theories, ideologies—Messick charged that we must take responsibility for each in interpreting scores. For someone with notable equanimity, Messick’s passage may be read as a cri de coeur. Because such force is absent in the Standards regarding validity theory, the 1999 edition is best read after reading Messick. Read in that order, the new emphasis on the five sources of evidence identified in Table 1 in place of criterion, content, and construct validity makes sense. As it turns out, it’s about both test design and test use.

Validation: The Significance of Fairness

In their call for a usable past for writing assessment, Brian Huot, Peggy O’Neill, and Cindy Moore remind us of the importance of contingency. Validity, they write, “is more than a range of statistical correlations” and the certainty they imply (511). If we are to ask and answer the really tough questions about validity in terms of fairness, as the new sixth edition of the Standards attempts to do, we must understand the trajectory of its historical past. Because a separation of scientific and social concerns is neither possible nor desirable, it is important

If the five previous editions of the Standards have been about validity, then the present edition is about fairness.
to remember that the concepts identified in the *Standards* are important in helping us make the best decisions on information generated through various measurement techniques. If there is a single instrumental lesson to be learned from the past, it is that interpretation and impact are historically contingent and inextricably bound.

If the five previous editions of the *Standards* have been about validity, then the present edition is about fairness. Interpreting the sixth edition within the framework of the origin and invention of measurement shown Table 1 allows us to understand just why the idea of validity—that “integrated evaluative judgment” that Messick advocated (13)—continues to be ordered up in the first place.

With its history in mind, review of the present version of the *Standards* is best begun with the expanded definition of fairness found in the glossary as the “use of test score interpretations for intended use(s) for individuals for all relevant subgroups. A test that is fair minimizes the construct-irrelevant variance associated with individual characteristics and testing contexts that otherwise would compromise the validity of scores for some individuals” (219). An armadillo of a sentence, it may be translated as follows: During measurement, some aspects of a construct may be distorted for some subgroups; when that distortion occurs, the validity of the scores is questionable. If individuals are visually impaired, for example, print and some electronic forms may disadvantage those who need magnification for reading. While commonsensical, this example has important implications for equity of opportunity. Throughout the new chapter, “Fairness in Testing,” we find that the committee—cochaired by Barbara S. Plake of the University of Nebraska and Laress L. Wise of Human Resources Research Organization—is more than willing to deal comprehensively with the topic, including technical properties of tests, reporting and use, elements impacting interpretation, and consequences of test use.

New to this edition is the introduction of the concept of universal design, an approach to assessment that strives to minimize construct distortion and maximize fairness through uniform access for all intended examinees. With origins in architectural planning associated with the Americans with Disabilities Act, universal design augments accessibility during the planning stages of an assessment (Ketterlin-Geller). Related to evidence-centered design as developed by Robert J. Mislevy and his colleagues, universal design seeks to strengthen the evidentiary argument that will eventually be used to justify the use of test scores by ensuring that barriers to access are anticipated and eliminated. Again, a very straightforward, commonsense principle has enormous
implications. Individuals with diverse gender, race/ethnicity, sexual orientation, culture, language, age, disability, and socioeconomic backgrounds (71)—the subgroups identified in the chapter, with the inclusion of sexual identity new and most welcome—must be considered in the assessment design process. When universal design is used, the possibility is increased that score comparability across groups can be identified and that differences in test scores among groups can be accompanied by cautionary statements. Recalling the introduction of regression analysis in the 1974 Standards as a technique to identify potential measurement bias, we can see the circle come round: during the planning stage of any assessment, universal design can be used as a way to leverage the validity of inferences from both individual and group test scores in the service of equitable treatment.

Of special interest to writing studies is the example of automated writing evaluation used in the discussion of response patterns. Along with test content, test context, and learning opportunity, the study of response processes—one of the five sources of validity evidence identified in the 1999 edition—may be used to identify threats to fairness. For instance, a scoring algorithm may not be sensitive to the diverse ways that individuals or groups express their ideas and, as such, may not produce scores valid for those individuals or groups (56). Further—and this is a groundbreaking recommendation of the new edition—“the precision of scores and validity of score interpretations resulting from automated scoring should be evaluated for all relevant subgroups of the intended population” (67). By logical extension, a system that does not provide such analysis should not be used when diverse groups of students are to be examined. Here is a use of the Standards that is relevant to all writing program administrators, one that applies to automated and human scoring equally.

To continue the review in terms of fairness as the first among equals in the validation process, we can now turn to the overarching standard—a new feature introduced in the 2014 edition:

All steps in the testing process, including test design, validation, development, administration, and scoring procedures, should be designed in such a manner as to minimize construct-irrelevant variance and to promise valid score interpretations for the intended uses for all examinees in the intended population. (63)

Perhaps it was not standardization itself that was the problem, as identified by the 1906 committee. Perhaps it was the orientation itself. The lesson from our usable past? Assessment should be, first, about fairness (Rawls).
Readers will have noticed the absence of the concept of reliability—the consistency of scores across replications of a measurement procedure—so far in this review. For far too long, reliability has been understood narrowly as the major precondition to validity. With its reconceptualization of the term reliability now presented as reliability/precision, the new edition of the Standards no longer implies that reliability is the chief sequential determinant of validity in complex assessments. While the Standards recognizes that steps toward greater flexibility in the measurement procedure increase the error associated with reliability (recall Concept 1), “some of these sacrifices in reliability/precision may reduce construct irrelevance or construct underrepresentation and thereby improve the validity of the intended interpretations of the scores” (36). Additionally, since validity associated with fairness is the aim of assessment, the authors also recommend that estimates of reliability/precision be provided for each relevant subgroup for the assessment. Specifically, claims of appropriate interreliability (recall Concept 2) need to be provided for each group in a given assessment to assure equitable treatment of those with diverse gender, race/ethnicity, sexual orientation, culture, language, age, disability, and socioeconomic backgrounds. In this view, reliability is no longer solely a precondition to validity; rather, reliability/precision is evidence of fairness.

And so it is that a new trinitarian model is formed: fairness, validity, and reliability/precision. To accompany this model, a new term is offered in the glossary: validation—“the process through which the validity of a proposed interpretation of test scores for their intended uses is investigated” (225). This new process-oriented model can then be refined through operations (from test development through rights of test users) and applications (from psychological testing through program assessment). Organized in thirteen chapters, the new edition of the Standards provides a comprehensive vision of what measurement at its best can and should be.

Writing Studies: A View of a Room
I have identified the affiliations of those who have contributed most to the Standards to demonstrate that its authors have overwhelmingly been university researchers, colleagues motivated by issues paramount to their own unique historical circumstances. Whatever visions of common good drove those who collected their paychecks in the academy over the past sixty years also motivated the authors of the six editions of the Standards. They are not the product of the testing industry. Each edition of the Standards should rightly
be considered a hallmark document with important conceptual advances in measurement featured in each edition.

While I have argued that the 2014 edition of the *Standards* is necessary reading for our field because of its innovative emphasis on fairness as a referential frame, I am not prepared to say it is sufficient for research in writing studies. My reservations are to be found in three inter-related fallacies.

*Platonism.* Philosophically, writing studies has never been especially drawn to anti-materialistic thought. Ours is a pavement-level profession. Many would indeed agree with Paul Miller that, while we have moved toward the incorporation of the material in conceptualizing our field, we still have not gone far enough. Informed by the philosophical heritage of our field, we can easily identify the implicit anti-materialism lurking through the *Standards.* As if conjured from *The Republic* itself, the language of the *Standards* is redolent of the metaphysical in its adoration of the abstract and its dismissal of the material. The discussion of the process of validation, key to the sixth edition, reflects the belief that there are essences that float just above the matter at hand. “Validation,” we are told, “involves attention to possible distortions in meaning arising from inadequate representations of the construct and also to aspects of measurement, such as test format, administration conditions, or language levels, that may materially limit or qualify the interpretation of test scores for various groups of test takers” (13). Assumed is a defined meaning, residing somewhere just beyond its measurement, that can be twisted out of shape in a heartbeat. Ours are shadows on a wall, thrown by a fire.

In writing studies, we take an alternative approach. We claim that there are constructs of writing congruent with specific curricular applications, such as the “WPA Outcomes Statement for First-Year Composition” (CWPA), that are, in turn, tailored to a particular setting. This definition and its local modification are all that exists. If things go sideways, it is unlikely that we would use the term *distortion* to describe the error. Rather, we would simply claim that knowledge of conventions received undue attention on the scoring rubric over, say, rhetorical knowledge, critical thinking, and composing processes. Ours are criterion-based assessments, welded to what students can do, not on aptitudes that hinge on what they should be able to do.

Platonism is also evident in the assumption that a test is fair when “it reflects the same construct(s) for all test takers, and scores from it have the
same meaning for all individuals in the intended population” (50). Considering the linguistic relativity that accompanies all language use (Gumperz and Levinson), the statement is remarkable in its implication that there resides a singular, interpretable meaning within a test that will result in scores that have the same meaning for all. In writing studies, we would look simply to elements in the test that might bias its design, anticipate iatrogenic consequences, and analyze the scores of each subgroup for differences that might indicate disparate impact. It is unlikely that anyone in our profession would claim that it would be possible for a test item to be understood in the same way by each individual or that the meaning of a score would be identical for everyone.

Value dualism. In her feminist theory of obligation, Nancy J. Hirschmann is quick to recognize the connection between Platonism and value dualism. In fact, she is reluctant to engage consent theory in the first place because its sense of duty is gendered and thus establishes epistemologies that are divisive. Such binary oppositions—Ann E. Berthoff referred to them as killer dichotomies—severely limit the Standards.

In the introduction itself, the authors note that “rigorous application of the Standards” in cases such as “instructor made tests that are used to evaluate student performance” is “generally not possible” (2). In the case of teachers developing classroom assessments, “it would be overreaching to expect that the standards of the educational and psychological testing field be followed” in their schools (3). Whatever is assessed in the classroom, the authors of the Standards appear to believe, is inferior because teachers lack rigor, intelligence, or both. Sadly, there is little room for teacher knowledge in this system. The killer dualism separating teachers from research is alarmingly apparent in a system that too often separates classroom instruction from educational measurement.

Once unleashed, dualisms run wild. It is unfathomable that the authors of the Standards treat public policy as an antagonist. “Tests used in policy settings may be subjected to intense and detailed scrutiny for political reasons,” the authors write. Yet because even empirical studies of high quality may be dismissed in policy debate, test users should “recognize that they cannot control the reactions of stakeholder groups” (207). Only gullibility would allow the authors to go out of their way to marginalize the very politicians they hope to influence. Only blindness would allow the authors to let anyone off the hook—designers, administrators, and politicians alike—when student welfare is under debate. Absent these causes of error, I cannot imagine why such a passage would not have been edited out; it is sure to alienate the very congressional aides who, it is hoped, would slip copies of the Standards into their bosses’ briefcases in
preparation for PARCC and Smarter Balance meetings.

Isomorphism. When Platonism is at play and value dualisms result, a compensatory strategy is to claim isomorphism of the system in terms of its utility. If the system is to be legitimate, the fallacy goes, its function needs to be uniform across time and circumstance. In reality, however, there is no sameness between the criteria and guidelines in the Standards and the test takers and test users who will be the targets of educational and psychological measurement. The example of special test preparation illustrates the disjuncture. “The integrity of test results,” we are advised, “should be maintained by striving to eliminate practices designed to raise test scores without improving performance on the construct or domain measured by the test” (213). In the case of placement tests, however, the authors acknowledge that commercial and local tests “do not measure prerequisite content” (187). If this is so, then idealism (never practice just to raise a test score) cannot be reconciled to the reality (endure the hazing of placement testing even though the scores reek of construct underrepresentation). Treated in such a way, it is no wonder that students may refuse to run the maze.

Measurement: As You Like It
If charged with identifying the origin of such fallacies within the Standards, my candidate would be Cronbach. As part of the project from 1952 to 1999, he defined a construct, the heart of the validity project, as “some postulated attribute of people, assumed to be reflected in test performance. In test validation the attribute about which we make statements in interpreting a test is a construct” (283). With philosopher of science Paul E. Meehl, Cronbach launched a project of quest in 1955 that continues to influence the ontology of the Standards.

That is not to say, of course, that latent Platonism informs all theories of measurement. A pragmatist, Michael T. Kane has defined his program of interpretation and use argument (IUA) as beneficial in interpreting specific claims based on particular test scores. “We have no direct access to Truth,” he has written, and so we can only “assume that sound methodology will promote what Popper (1962) calls ‘verisimilitude’ and Lakatos (1970) refers to as a ‘progressive research program,’ but making Truth the centerpiece of our conception of validity is not likely to be helpful” (“Validation as a Pragmatic” 121). He was a member of the joint committee authoring the 2014 Standards, and we can only wish that more of his voice had been present in the final text. Within our field, we can look to Asao B. Inoue, who has proposed linking fairness to the very
definition of validity itself through *racial validity*, defined as "an argument that explains the degree to which empirical evidence of racial formations around our assessments and the theoretical frameworks that account for racial formations support the adequacy and appropriateness of inferences and actions made from the assessments" (110). Conceptual scholarship concerning race and identity formation are continued in *Race and Writing Assessment*, edited by Inoue and Mya Poe.

Yet even with these exemplar works in hand, we must acknowledge that sufficiency is a high bar to set. We will not achieve it in writing studies until we have centralized the necessity of replicable, aggregable, and data (RAD) supported research in our field and trained a generation of graduate students through programs associated with the Doctoral Consortium in Rhetoric and Composition—an absence that Benjamin Miller has documented in his study of 2,711 dissertation abstracts from our field. Creating a core curriculum with specialized research electives in theory building through quantitative and qualitative methods—not courses borrowed from the math or psychology departments but our own courses tailored to problems and techniques in our field—will yield the progressive program of RAD research advocated by Richard H. Haswell and made possible, in part, by the *Standards*.

Readers will note that we have come full circle to justify my initial claim that the system of educational measurement, as embodied in the *Standards*, is relevant for writing studies because it informs our research paradigms including, but not limited to, writing assessment. The value dualism that has divided writing assessment from research, as Huot has noted, has "obscured the essential purpose of assessment as research and inquiry" (148). Deconstructing the dualism allows us to see the RAD commonality among many of the most frequently cited authors in *College Composition and Communication* from 1987 to 2011. While disciplinary density is hard to tie down, as Miller has documented in his study of citation patterns, the unifying concepts of measurement expressed in the *Standards* provide a way to understand that which binds us together as a field.

Since reading the *Standards* will force you to examine your beliefs about truth and method, set aside the two long days it will take to read and react to the 230 pages of two-column text. Go somewhere quiet and bring your knowledge of everything you know about writing studies with you. Confirm what you know to be true, open up some space for new possibilities, and rage at the parts where the light seems to dim. After you have put yourself through your paces, whatever remains will be something like the truth.
Acknowledgments

I would like to thank Richard H. Haswell, Irvin R. Katz, and Frances M. Ward for their perceptive readings of this review. I would also like to thank the Educational Testing Service for access to archives in Carl Campbell Brigham Library. The exposition, interpretation, and critique are mine.

Works Cited


——. “Validation.” Brennan 17–64.


Norbert Elliot

Norbert Elliot is professor emeritus of English at New Jersey Institute of Technology. He is author, most recently, of Henry Chauncey: An American Life. He has been a member of NCTE/CCCC since 2001.