4
THEORY UNDER CONSTRUCTION

The work of developing theoretical principles specifically for writing assessment has begun in the last decade, but it remains in nascent form. Influential texts in composition studies such as those by scholars in the CCCC Committee on Assessment and by Brian Huot suggest principles and procedures for contextually- and rhetorically-aware assessment; however, neither presents a fully-articulated theory of writing assessment—the former because of the project's rhetorical purpose, the latter because the author considers such a move premature.[30] While both of these texts have been influenced by positivist educational measurement thought, both also develop positions which correspond more nearly with contemporary literacy theory and instruction than do the best assessment methods developed from within an educational measurement tradition.

For some scholars, Huot included, the attachment to educational measurement theory goes beyond the historical predisposition and determinism that appears in many texts, and is, rather, indicative of a movement within composition studies to co-opt educational measurement theory. Unlike White's position that compositionists should accept educational measurement theory and apply it as that field does, these co-opters work to appropriate the theory and adjust and apply it within parameters more conducive to thinking in composition studies. Huot's work, in particular, lauds contemporary validity theory. Rather than accept reliability, however, he argues that current thinking and procedures in writing assessment make that principle moot.

But I do not think it is that easy. While there is much to celebrate in the theoretical changes in validity, reliability remains a key concept in educational measurement theory and thus remains a problem. This is not to say that educational assessment theorists are satisfied with the dominant approach, however. The work of Pamela A. Moss and others focusing on complex performance assessments challenges reliability as it appears in the dominant paradigm, but their alternatives have not reached the establishment, and the limitations imposed by traditional educational measurement theory on writing assessment remain intact. Reliability still serves as a limiting condition on validity, and consequently on large-scale assessment as it is practiced.

Co-optation is not our only option. Within composition studies, some scholars have posed direct challenges to educational measurement theory, particularly to reliability. Richard Haswell's work on categorization theory, discussed briefly in the last chapter and more fully in this one, provides theoretical justification for the expert reader model used at Washington State University. Limitations of Haswell's theory, however, make it difficult to apply beyond placement assessments. Other approaches, including Bob Broad's use of hermeneutics and inquiry theory, have suggested different changes, though these theoretical efforts tend to remain speculative or of limited application.

Whatever the limitations, taken together, these changes in and challenges to educational measurement theory indicate dissatisfaction with the principles as they exist, and they suggest a paradigm shift in progress. However, the traditional principles remain in force, despite these challenges. The paradigm has not shifted. Yet these theoretical alternatives suggest a kind of movement and indicate a willingness—and perhaps even a readiness—to try a different framework, certainly within composition studies and possibly within educational assessment.

THE PARTY LINE

In his response to Elbow and White in *Composition in the 21st Century*, Brian Huot argues that composition lacks "a Theory of Writing Assessment [*sic*]" (1996a, 115). Without one, he claims, assessment practices will not reflect the theories and practices of writing and its learning either now or in the future. In the last decade, we have seen some scholarship within composition studies that begins the work of developing theoretical principles sensitive to a contextual literacy paradigm. The most influential of these take educational measurement theory as a foundation or starting point. For the most part, this scholarship relies on contemporary validity theory *à la* Samuel Messick, which has been generally accepted in educational measurement theory and which I will return to in a moment, and rejects or argues around reliability. However, the mainstream of contemporary educational measurement theory still treats reliability as a precondition for validity, and challenges within educational measurement have not been successful to-date. Still, the movement to co-opt educational measurement theory is strong in composition studies.

Published in 1995, "Writing Assessment: A Position Statement" developed by the CCCC Committee on Assessment does not specifically claim theoretical status for itself, but it does offer a set of principles for sound practice, and given that the Statement was developed under

the sponsorship of CCCC, it speaks with the voice of authority, at least within composition circles. I address this document here because as a Position Statement, it articulates best practices and carries the force (at least potential) of theory combined with practice. The Committee begins with a foundational premise: "the primary purpose of the specific assessment should govern its design, its implementation, and the generation and dissemination of its results" (1995, 431). In order to direct assessment procedures, the primary purpose would need to be articulated—presumably by those initiating the assessment procedure—and then from that aim, practice could follow. From this basis in purposeful assessment, the committee then offers a set of practical guidelines: "Assessments of written literacy should be designed and evaluated by well-informed current or future teachers of the students being assessed, for purposes clearly understood by all the participants; should elicit from student writers a variety of pieces, preferably over a period of time; should encourage and reinforce good teaching practices; and should be solidly grounded in the latest research on language learning" (1995, 431). These guidelines argue for the primacy of pedagogy, implicitly claiming that assessment of writing serves the instruction of writing first and foremost. Taken together with the foundational principle above, the Committee asserts that writing pedagogy provides the context for writing assessment and should therefore guide all aspects of assessment from design to dissemination.

The Committee elaborates on this context through a set of ten assumptions that the members claim should provide the basis for all writing assessments, and which thus serve as the foundation for their position statement. The first four assumptions reflect the values of what I have called the contextual paradigm of literacy:

1. language is always learned and used most effectively in environments where it accomplishes something the user wants to accomplish for particular listeners or readers within that environment;
2. language is by definition social;
3. reading—and thus, evaluation, since it is a variety of reading—is as socially contextualized as all other forms of language use; and
4. any individual's writing "ability" is a sum of a variety of skills employed in a diversity of contexts, and individual ability fluctuates unevenly among these varieties. (1995, 431–32)[31]

Taken as a group, these principles require that writing assessment be grounded in the same context as the writing itself—however that particular

context is determined—and that the evaluators perform the assessment with that context in mind.

The next five assumptions appear to be the result of some well- and hard-learned lessons about the effects assessment has had on writing instruction:

5. writing assessment is useful primarily as a means of improving learning;
6. assessment tends to drive pedagogy;
7. standardized tests, usually developed by large testing organizations, tend to be for accountability purposes, and when used to make statements about student learning, misrepresent disproportionally the skills and abilities of students of color;
8. the means used to test students' writing ability shapes what they, too, consider writing to be; and
9. financial resources available for designing and implementing assessment instruments should be used for that purpose and not to pay for assessment instruments outside the context within which they are used. (1995, 432–33)

These assumptions serve, in part, as a codified response to the abuses that writing instruction has suffered under the positivist paradigm of writing assessment. The first two of these specifically address the connections between pedagogy and assessment, contending that testing should reflect the best of classroom practice and that the results of the tests should be useful within the classroom. The last three speak specifically to the effects of objective testing, including the socio-political ramifications, the power of testing to define what is valued, and the delegitimating consequences of outside testing.

The final assumption defines an overarching trajectory for writing assessment:

10. there is a large and growing body of research on language learning, language use, and language assessment that must be used to improve assessment on a systematic and regular basis. (1995, 433)

The Committee here links scholarship and assessment, arguing that research in the field of composition studies is not only relevant, but also essential to the improvement of assessment practices. The pedagogical emphasis of the rest of the assumptions makes sense considering that teaching has been subordinated to testing for a long time; research has not. This assumption tacitly legitimates the first nine pedagogically oriented assumptions by claiming that research supports these presuppositions. For those within composition studies, this assumption is rather

redundant: current research is already reflected in the earlier assumptions. The audience and purpose for this document, however, make it necessary to lay out this principle.

This audience is explained at two points. First, the immediate audience is outlined in the introductory section, which describes the document's origins and process of development. It initially developed as a response to members of CCCC who wanted a document "that would help them explain writing assessment to colleagues and administrators and secure the best assessment options for students" (1995, 430). Those "colleagues and administrators" presumably would not be knowledgeable about research in the field, and this statement would explain some of the most basic principles. In this sense, it can serve a defensive purpose: explaining the principles of assessment to those who would question assessment practices developed in accordance with this statement, or worse yet, mandate something unprincipled.

The second point at which the Committee describes the audience assumes a more assertive posture. The final third of the statement enumerates the rights and responsibilities of the primary stakeholders; although, the Committee only uses that term indirectly. This list of who "should" do what—students, faculty, administrators and higher education governing boards, and legislators—bears a distinct resemblance to White's stakeholder list.[32] However, instead of asking what these constituencies want, the Committee diagrams their expected participation in writing assessment. This part of the document takes a more directive tone, explaining what the members of each group are accountable for, as well as what they can expect.

This document is important in large part for the way in which it claims authority for the practice of writing assessment. Unlike constructivist evaluation methods, which make the concerns of each stakeholder the responsibility of all the participants, and unlike the numerous assessment situations in which composition professionals—alone or together with students—are held accountable for writing assessment outcomes, this position statement holds each constituent responsible for its own informed, ethical, and appropriate participation in the assessment process. In part because of its pragmatic emphasis and its relatively broad audience, this document downplays the specific theoretical principles that bolster its claim to authority. A specifically theoretical text would address itself to members within the discipline and would tend to rely on a higher level of abstraction. In a document intended to explicate policy, such abstraction is hardly welcome.

The theoretical principles of writing assessment that underwrite this document, however, are neither clear nor well developed in the literature to-date. As my discussion of contemporary assessment models in the last chapter indicates, the traditional principles of educational measurement theory cannot easily account for notions such as contextualized expertise or assessment as an ongoing and evolving process. Yet composition studies does not have well developed and accepted alternative theoretical principles in place. This position statement presents assumptions and guidelines based on the best current thought on writing and learning to write, where "best" has been determined by research trends in the disciplinary literature—thus the gesture of the tenth assumption. This thought, however, has yet to be gathered in a systematic or developed manner.

Brian Huot does some of this developmental theoretical work in his recent book, *(Re)Articulating Writing Assessment for Teaching and Learning* (2002), which frames writing assessment as a field and examines that field for ways in which it might be reconstructed more productively. The specifically theoretical part of Huot's venture appears in chapter four, which is a revised and expanded version of "Toward a New Theory of Writing Assessment" (1996b), an essay that originally appeared in *College Composition and Communication*. Huot's thinking about writing assessment theory evolves from two directions: contemporary writing assessment practices that circumvent positivist epistemology and state-of-the-art validity theory developed by measurement scholars. Writing assessment, he argues, has been controlled by the measurement community, and he points out that the measurement and composition communities have distinct theoretical differences and share an inability to communicate across them. He begins construction from the composition side of the gap by summarizing some of the ways in which contemporary validity theory applies to some contemporary writing assessment practices. Thus, by sifting through the most promising writing assessment practices, Huot begins the important—and overdue—work of outlining principles that can be used to theorize writing assessment.

Huot focuses his analysis of practices within composition studies on the ways that some contemporary assessment procedures, such as the expert reader model, are grounded in specific institutional contexts that define the purpose of their assessments. Examining models such as those presented by Smith (1993) and Haswell (2001c), he concludes that cutting-edge procedures such as these are sound specifically because teachers who thoroughly understand the curriculum make the placement decisions, and thus that context is inherent and necessary to the decision

being made. He also examines the discussions on the "portnet" listserv presented by Michael Allen in "Valuing Differences" (1995), and concludes that a general knowledge of the system in which a portfolio is to be assessed is sufficient to generate agreement about assessment when the local context is foregrounded. His final example, "Portfolio Negotiations: Acts in Speech," by Russel K. Durst, Marjorie Roemer, and Lucille M. Schultz (1994), describes an exit portfolio assessment program in which "trios" of teachers in conversation make final judgments, a practice I return to in chapter seven. Huot argues that all of these methods "share assumptions about the importance of situating assessment methods and rater judgment within a particular rhetorical, linguistic, and pedagogical context" (2002, 98).

Huot's conclusions about the contextuality of assessment coincide with the conclusions of contemporary literacy scholarship, but he also argues that contemporary validity theory, as it is evolving among measurement scholars, provides a sufficient foundation for principles of writing assessment. Traditionally, as noted earlier, validity has meant that "the assessment measures what it purports to measure" (Huot 2002, 87). In classical measurement theory, validity relies in significant part on an objectivist understanding of writing ability as a fixed, acontextual property that resides entirely within the text at hand. Current validity theory, however, makes both the context and the use of a test an explicit part of validity.[33] Huot quotes Samuel Messick, a leading validity scholar whose definitional work has been instrumental in revising this principle and whose work I will return to shortly: "validity is 'an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores or other modes of assessment'" (Huot 2002, 93; Messick 1989a, 5).[34] Contemporary validity theory, that is, claims that validity as a criterion for judging assessment practices is not meaningful unless it includes a sound theoretical base for the object and means of assessment; appropriate consequences from the results of the testing; and empirical evidence generated by the testing of both these qualities. The implicit argument in Huot's review of contemporary validity theory is that compositionists would find much of value in the work of the measurement community if we could or would understand it.

While he clearly favors contemporary validity theory, Huot challenges reliability as a defining concept in writing assessment. In traditional testing theory, reliability is a necessary condition for a valid test; if a test can not be scored consistently, it can not be valid. Because validity presupposes

reliability, efforts in writing assessment historically have focused on the latter. Huot suggests that the positivist environment—which focused on universality and generalizability—encourages evaluators to strip the context from pieces of writing used in assessment. This context-stripping, he argues, makes reliable assessments impossible because readers have little basis for judgment. Huot points out that norming procedures establish reliability not through developing consensus, but through rebuilding a context, and he argues that recent assessment methods, particularly the expert-reader model, have been so successful because they skip the steps in which context is stripped and rebuilt, and instead leave the context intact.

In addition, Huot discredits the notion that reliability ensures fairness to students. He points out that "reliability indicates only how consistent an assessment is," and that consistency is only one part of fairness, not the sum-total (2002, 87–88). Fairness, according to Huot, should also include information about the basis for evaluation. He argues that "[t]ranslating 'reliability' into 'fairness' is not only inaccurate, it is dangerous, because it equates statistical consistency of the judgments being made with their value" (2002, 88). His claims here echo earlier work done with Michael M. Williamson (2000) that challenges White's assertion that fairness is another way to understand reliability. In their examination of the relationship between ethics and assessment, they argue that "White's claim that a test *must* be reliable to be fair . . . frames a technical issue in ethical terms" (2000, 194). Reliability, in its psychometric home, is a criterion for measuring something more akin to equality than to fairness, and they point out that "fairness involves reasonableness, not equality" (2000, 194). To equate fairness with reliability is to elide issues of power and access, and to deny the complexity of the entire assessment. In *(Re)Articulating*, Huot concludes that the best of contemporary assessment procedures and the theoretical principles he outlines either bypass inter-rater reliability or make it moot in the face of contextually-bound evaluations (2002, 98).

The result of Huot's exploration is a set of five principles extrapolated from the practices he discusses: writing assessment should be "site-based," "locally-controlled," "context-sensitive," "rhetorically-based," and "accessible" in its entirety to those being evaluated (2002, 105). The first three are all variations on the theme of contextuality: assessment should arise from local need, remain within local control, and be sensitive to local issues. By "rhetorically-based," Huot means that assessments and the pedagogical situations that produce them should be grounded in current thought in composition studies, and particularly in literacy research. The

final principle addresses the ideal of fairness through access to information. According to Huot, all of these principles begin with the notion of context and will require new procedures for validation, which will likely include qualitative and ethnographic methods.

Huot's reliance on contemporary validity theory seems somewhat incongruous with his self-defined project to "explore our [compositionists'] ability to construct a theory of writing assessment based upon our understandings about the nature of language, written communication, and its teaching" (2002, 94). However, his work, first published in 1996 and revised in 2002, is the first in over a decade that explicitly attempts to develop principles for all writing assessment situations based on current thinking about literacy (I will examine some more limited attempts later in this chapter). More importantly, Huot's principles roughly parallel those offered by the CCCC Committee on Assessment, indicating some level of agreement among influential scholars in writing assessment within this discipline about appropriate or valuable writing assessment practices, if not about the specific principles supporting those practices. Thus, both the Committee's and Huot's texts begin the work of drawing together the practice of writing assessment and the principles of the contextual paradigm of literacy.

ON THEIR TERMS

Huot and the Committee are hardly the only compositionists who favor co-opting educational measurement theory for the purpose of writing assessment. We can see variations on Huot's argument in the work of a range of scholars, including White (1994b), Hamp-Lyons and Condon (2000), Yancey (1999), and Scharton (1996), among others. The specifics of the argument vary, but the acceptance of educational measurement theory is more than bowing to those with power, and historical precedent only partially explains the persistent connection. Instead, like Huot, advocates of co-optation are claiming master's tools, arguing implicitly that compositionists can use them, too—if not to tear down the master's house, at least to do some much-needed renovations.

This is a risky proposition. Compositionists do need principles for determining the value of an assessment procedure that develop from theories of composition and literacy learning. But if successful co-optation is possible—and I am not convinced it is—we will need to address the relationship between the measurement theory that the terms we are "borrowing" represent and writing assessment as it is currently practiced. If we continue to use educational measurement theory to legitimate assessment practices,

we allow measurement theorists to determine the value of writing and writing instruction. And where these theories incorporate objectivist ideals, we perpetuate the distance between writing theory and pedagogy on the one hand and writing assessment on the other.

The position statement developed by the CCCC Committee on Assessment does perhaps the best job distancing writing assessment from the criteria of educational measurement theory. The document makes reference to these terms, but the Committee does not use them specifically to ground the principles they outline. They do argue that faculty have the responsibility to familiarize themselves with "the norming, reliability, and validity standards employed by internal and external test-makers" (435), but these criteria seem to be subordinate to the principles developed from composition theory and practice. The relationship, however, is not clear, nor is explicating the relationship within the scope of this position statement. The political nature of this document encourages a stance that foregrounds composition's principles. Because of its form and purpose, it has been treated less like theory and more like a public declaration, and thus used more to justify practices to those outside of composition studies than to direct practices within it. So while this document seems to do a better job grounding assessment practices specifically in composition theory than Huot's, it is ultimately less effective in developing an assessment theory.

Huot's book is influential, and his position is the most fully articulated of those I have discussed. It is also the most conflicted. He argues that cutting-edge validity theory corresponds to cutting-edge theory about writing, and that reliability is made moot by context. But while he references neither "reliability" nor "validity" in the figure that outlines his principles, he seems ambiguous about the relationship between writing assessment theory born of composition theory and the principles of measurement theory. Even as he discusses the promise of contemporary validity theory, he acknowledges that the technicalities of concepts such as validity and reliability have alienated most compositionists. He tries to explain contemporary validity theory in terms that compositionists will understand, and eventually claims that this theory supports "a new theoretical umbrella" which gathers together the principles at work in the best of contemporary assessment methods (2002, 104). In this metaphor, which he uses several times, measurement theory holds the umbrella. Or, framed in a more evenly distributed way, Huot seems to be trying to ground a theory of writing assessment by placing one leg in composition theory and the other in measurement theory. The implication is that composition theory

by itself does not or cannot—it is not clear which—provide sufficient criteria for judging methods for assessing writing.

Other scholars make similar arguments. White, for example, has been arguing for years that we must pay attention to educational measurement theory or the psychometricians will do our assessing for us. William Condon, in both his research with Liz Hamp-Lyons (Hamp-Lyons and Condon 2000) and in his work on the assessment system at Washington State University (Leonhardy and Condon 2001), treats validity and reliability as a kind of point zero—though he does tend to characterize these criteria in much the same way Huot does. Kathleen Blake Yancey, in her history of writing assessment (1999), describes the phases of writing assessment in composition studies in large part in terms of the influence of reliability and validity on the shape of assessments in the field—though she is not clearly advocating the use of these principles.

In an essay published the year after the original version of Huot's theoretical argument, Huot and Williamson (1997) make the case that problems in educational assessment cannot be solved at the theoretical level, or perhaps more accurately, that theory alone will not solve the problems. They review the literature from educational measurement, focusing on the work of Moss in particular, for the ways in which theoretical principles are changing. They conclude, however, "that oftentimes issues of power rather than theory drive important assessment decisions" (1997, 44). I agree, but I would add that theory and power are historically and rhetorically tied to one another in the case of writing assessment, as I discuss in the first chapter. Huot and Williamson work to separate these issues primarily to bring the problem of power to the attention of compositionists. In the process, however, they continue to refer to educational measurement theory and theorists for the "theory" part of their equation, leaving the theoretical power squarely in the hands of educational measurement.

Huot claims that "[f]ew important or long lasting changes can occur in the way we assess student writing outside of the classroom unless we attempt to change the theory which drives our practices and attitudes toward assessment" (2002, 94). Yet he implicitly argues that a significant part of that change comes in the form of compositionists' recognition and understanding of contemporary validity theory as espoused by Messick and Moss. I am not convinced that this constitutes a sufficiently significant change, given the baggage that validity and reliability carry and their political influence in the practice of large-scale assessment. Granted, there is value in co-opting educational measurement theory.

The principles have a long and established history, which gives them currency in the influential market of testing. This strategy might work, but co-optation has its limitations—in this case, mostly in the form assessments take in the "real world," as we can see by the use of state-level K-12 writing assessments and many of our own college-level large-scale placement and exit exam procedures. As long as compositionists continue to rely on validity and reliability, we leave ourselves open to charges of not using the principles correctly and demands that we reconfigure our large-scale assessments to meet the established requirements of the powers-that-be.

CHALLENGES WITHIN EDUCATION

Thus far, I have constructed an us/them narrative, a story about how educational measurement theory has dominated writing assessment practices. But just as theories of writing are not monolithic in composition studies, neither are theories of educational assessment in that field. While the history of educational measurement has had decidedly objectivist foundations, not all current theory in that discipline continues the tradition, and there are a number of examples of theorists within the field who are challenging key concepts in the theory. In addition to Messick's work in validity theory, those working in complex performance assessment in particular demonstrate the ways in which theorists within that field are dissatisfied with classical educational measurement theory. This work is important and may well help us work through our own difficulties with writing assessment.

Samuel Messick—considered perhaps the preeminent validity scholar in educational measurement theory at least through his death in 1998—is the author of the "Validity" article in the third edition of *Educational Measurement*, the most recent definitive statement of accepted theory in educational measurement (Linn 1989). His is the standard definition of validity: "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores or other modes of assessment" (Messick 1989b, 13), the definition that Huot cites as exemplifying the complexity and value of validity for writing assessment. Among Messick's contributions was the move of validity away from its status as a characteristic of a test, a property of the instrument, and toward an emphasis on the uses of the results of tests as a determination of validity. This shift means that no assessment procedure can be determined "valid" once and for all; determinations would be necessary for each use

of an assessment instrument. The complexity of validity as Messick defines it is more in keeping with the kinds of complexity that appear in writing assessment and provides solid justification for using complex performances in testing situations.

Performance assessment deals with the evaluation of complex tasks: open-ended questions, portfolios, hands-on experiments, and the like. These kinds of tests produce results that are not easily reduced to statistically reliable numbers. In the late 1980s and early 1990s, articles on performance-based assessment began appearing in education journals, among them Robert L. Linn, Eva L. Baker and Stephen B. Dunbar's "Complex, Performance-Based Assessment: Expectations and Validation Criteria" (1991). Linn, Baker, and Dunbar analyze existing educational measurement theory for the ways in which it in appropriately delimits performance assessments. They conclude that "[r]eliability has too often been overemphasized at the expense of validity; validity itself has been viewed too narrowly" (1991, 16), and thus, at least as they are traditionally conceived, these principles are too limited to provide a strong set of criteria for the level of complexity in performance assessments. The authors' purpose is to provide an expanded understanding of validity for evaluating assessments that is sensitive to the needs of more complex instruments. In this vein, they propose eight criteria: consequences, fairness, transfer and generalizability, cognitive complexity, content quality, content coverage, meaningfulness, and cost and efficiency.

These criteria are worth defining briefly for the ways in which they complicate the concept of validity. Drawing on Messick's work, "consequences" focuses on the effect of the assessment on pedagogy in particular, and the authors make clear that this criterion is paramount. "Fairness," while more complex here than in White's configuration, is primarily focused on issues of bias and equity in terms of race and class, and the authors argue that "[t]he training and calibrating of raters is critical" in eliminating evaluators' biases (1991, 18), reinforcing the historical connection between fairness and objectivity. Reliability is largely subsumed under the criterion of "transfer and generalizability." Linn, Baker, and Dunbar review the research to-date that demonstrates that performance is "highly task dependent" and performance-based assessments offer only a "limited degree of generalizability" (1991, 19). They suggest that to meet this criterion, students might be assigned more tasks or evaluators could use sampling for statistical purposes, but their argument does not provide clear direction for how educators and assessors would deal with this problem.

The authors spend much less time on the remaining criteria. "Cognitive complexity" examines the level of higher-order thinking skills assessed, and "content quality" analyzes the relationship between the content of the assessment and current best thinking in the relevant field. Analyzing content quality, they argue, requires subject matter experts, the only place in the assessment process where Linn, Baker, and Dunbar explicitly include such experts. Similarly, "content coverage" focuses on the comprehensiveness or scope of the assessment and serves to encourage breadth in the curriculum, given that what is not tested often is not taught or learned. The authors do not include experts here, but more interestingly, they argue that traditional testing may have an advantage over performance assessments in this area because of the ways that assessment drives the curriculum; performance assessments tend to produce depth rather than breadth. "Meaningfulness," a term I will return to shortly, here deals simply with the relevance of the task(s) to students, and "cost and efficiency" deals with the practical realities of testing. Taken together, these criteria offer an expanded understanding of validity that, the authors argue, "more adequately reflect theoretical concepts of validity" (1991, 20).

These criteria did not move far from traditional educational measurement criteria, at least as we currently understand them, but they were a bit of a stretch at the time. In 1991, Messick's claim for the importance of evaluating the consequences of the assessment decision and for attaching validity to that decision rather than to the test itself was still a fairly new concept, though sufficiently established to be included in *Educational Measurement* (Messick 1989b). Linn, Baker, and Dunbar acknowledge that they are only providing an expanded practical application of existing validity theory, but they do something a bit more radical with reliability. Buried in this piece is a fairly quiet challenge to that criterion: while the authors agree that "efficiency, reliability, and comparability are important issues that cannot be ignored in new forms of assessment," they also point out that these criteria "should not be the only, or even the primary, criteria in judging the quality and usefulness of an assessment" (1991, 16). Given that reliability had been (and still is in most assessments) a precondition for validity, this attempt to shift it to secondary status is noteworthy. The authors stop short, however, of repudiating these principles, and instead argue that "[n]onetheless, they [efficiency, reliability, and comparability] are issues that will require attention and careful design of procedures to assure that acceptable levels are achieved for the particular purposes of an assessment" (1991, 16). While they do not insist on reliable assessments, they do claim that the criterion will continue to

serve a useful—and limiting—purpose. This statement leaves the authors attached to educational measurement theory.

Pamela A. Moss's challenges to reliability are more direct. Like Linn, Baker, and Dunbar, Moss comes at her challenges to reliability through both performance assessment and through contemporary work in validity theory. In "Shifting Conceptions of Validity," she points out that current consensus dictates the primacy of construct validity and the need for consideration of the consequences of any given assessment (1992, 230). However, she also notes that "the practice of validity research typically has not done justice to the modern views of validity articulated by Cronbach and Messick. In fact, researchers still tend to rely on traditional sources of evidence—such as, evidence about content representativeness, internal consistency (and reliability), and correlations with alternative measures—which [many validity scholars working in performance assessment] consider insufficient for the evaluation of performance assessments" (1992, 245). Her point is that while the theory has progressed, the practical application of it has not. Moss's primary concern is that the demands for reliability and generalizability, which are excessively difficult to meet in complex performance assessments, dominate the field of assessment so that standardized testing has been and continues to be privileged over more complex instruments that often give more useful results.

In "Can There Be Validity Without Reliability?" (1994), Moss more fully explicates her hermeneutic alternatives to standardized assessment practices and thus deals more directly with the problem of reliability. Hermeneutics, she argues, provides an interpretation-based theoretical approach to the process of evaluation that stands in contrast to the quantification and consistency demanded by the psychometric principle of reliability. She posits dialogue toward consensus among those with expertise as an appropriate alternative to interchangeable, objectively oriented, and quantifiable assessments. In a hermeneutic assessment, experts would all read and interpret the assessment instrument, and then would discuss the results as necessary. This, she points out, is what educators already do with some of our more important decisions, such as hiring and tenure.

Moss is careful to point out, however, that she is not interested in overturning reliability as an assessment principle, but rather in offering another theoretical possibility (1994, 5), a position she reiterates two years later in "Enlarging the Dialogue in Educational Measurement: Voices from Interpretive Research Traditions" (1996). Here, she argues that she does not want "to overturn the foundation of educational measurement, but rather to extend it by locating conventional theory and practice within

a broader field of possibilities" (1996, 20). Using the example of teaching portfolios, she argues that a conventional psychometric evaluation, where readers assess each entry independently and without knowledge of other parts of the portfolio or the particulars of the teacher, results in a low-quality evaluation with consequences made questionable by the procedure. She asks whether "expert readers, familiar with the candidate's capabilities across multiple performances, may not result in a more valid and fair decision" (1996, 25). Here, Moss clearly challenges the positivist underpinnings of assessment practice: evaluating the parts separately does not, according to Moss, result in a clearer or more accurate picture than evaluating the whole all together.

Much of the work in performance assessment in the field of education, including much of Moss's, involves the assessment of teachers. Ginette Delandshere and Anthony R. Petrosky (1998) take up this process in "Assessment of Complex Performances: Limitations of Key Measurement Assumptions," where they describe their development of a certification procedure for the National Board for Professional Teaching Standards in Early Adolescence/English Language Arts. For certification, at the time of Delandshere and Petrosky's work, teachers were required to submit a portfolio documenting three teaching activities and to participate in "an assessment center" where they were interviewed, analyzed their teaching, and wrote essays on educational topics in a testing setting (Delandshere and Petrosky 1998, 14). The complexity and volume of the material made standard evaluation procedures difficult at best and, more importantly, made the standard numerical results of limited value.

Delandshere and Petrosky focus their work in this piece on the tension between an evaluation system based in measurement and one based in interpretation. The procedure they developed for National Board certification includes the use of interpretive summaries written by judges of the teachers' work, based on those judges' perceptions of teacher performance. Delandshere and Petrosky chose this approach over the traditional measurement approach, which would result in numeric scores, because they believe that the measurement approach produced results too limited to be of use. They argue that "[r]educing performances to a set of scores and generic feedback satisfies the needs of a certification decision but falls short of providing useful representations and analyses of actual teaching performances" (1998, 16). That is, while it is possible to meet the needs of statistical reliability and validity, the results are not particularly helpful.

The influence of measurement, however, is significant. Traditionally, educational assessment has resulted in numbers that can be generalized

and compared. Interpretive results complicate notions of generalizability, raising questions about the need for and use of generalizations from test results in complex situations. Delandshere and Petrosky point out that, because in complex performance assessments the tasks involved are constructed and context-bound, they are not generalizable in the same way that sampled tasks are. They acknowledge that educational assessment experts still want to be able to generalize, but they argue that there is an inherent "trade-off . . . between the universe of generalization and the complexity of the construct being assessed" (1998, 20). Generalization, a key principle in educational measurement theory and the driving force behind reliability, they argue, may be incompatible with useful assessments of complex performances.

Because of the political and practical demands involved in developing procedures for a real high-stakes, national-level assessment, Delandshere and Petrosky could not walk away from generalization in their project, so to answer the need for reliability, they constructed rubrics that would translate the complex evaluative statements developed by the judges into scores. However, they "found the process contrived, somewhat arbitrary, and overly reductionist" (1998, 21). Not surprisingly, they also found the results of limited use, and they conclude that numerical results "are poor representations of complex events" (1998, 21). Of course, this is not news to the composition community. Peter Elbow, for example, has argued for some time that numbers do not provide students with any useful information about their writing (1993; 1994), and the work on how best to respond effectively to student writing far outweighs the work on how to put a grade on it.

The pressure for numeric results, however, is systemic. Echoing Moss's concerns about the way validity theory shows up in practice, Delandshere and Petrosky point out, as an aside, "that when presented theoretically, discussions of validity include both quantitative and qualitative summaries. . . . When, on the other hand, the discussions turn to concrete examples of evidence (e.g., correlation, factor analysis, test variance) to be used in support of validity arguments, the same authors almost always refer to numerical scores or ratings" (1998, 16). Even contemporary validity scholars such as Messick, they claim, return to numeric examples to provide their evidence. Delandshere and Petrosky suggest that this reliance on numeric evidence continues because work outside the measurement tradition—as opposed to the interpretative tradition—in educational assessment has not been developed. Without the development of practices and principles, interpretative approaches to assessment tend to

founder and are overlooked in favor of established quantifiable methods. This certainly has been the experience of many of those of us in large-scale writing assessment, where we must provide numeric data to support our practices, whether we find the numeric data useful or not.

It has also been the experience of those in education. For example, the National Board revised its assessment for Early Adolescence/English Language Arts in 2002, the assessment Delandshere and Petrosky worked on. This "Next Generation" of assessment still includes the portfolio in largely the same form as the previous year; the directions ask for four entries: one based on student work, two on videotaped class sessions, and one on documentation of teaching accomplishments outside of the classroom (National Board for Professional Teaching Standards 2003a, 9–13). The assessment center questions, however, are different, and the differences are not what I would call positive, at least from the perspective of someone trying to gain an understanding of how a teacher teaches. The "Original Format" questions asked teachers to answer four 90–minute prompts based on material provided in advance, including reading lists, professional articles, and student writing. These prompts asked teachers to develop course materials, prepare for writing instruction, analyze student language use, and provide guidelines for text selection (National Board for Professional Teaching Standards 2003c), all activities a teacher would need to be able to do well. The "Next Generation" questions, on the other hand, ask teachers to answer six 30–minute prompts based on poetry, short text selections, and student writing, none of which are provided in advanced. In these short impromptus, teachers are asked to analyze literature, to discuss themes in an imaginative work, and to analyze non-fiction prose for audience and purpose. They are also asked to develop teaching strategies for correcting reading misapprehensions, continuing language development, and correcting weaknesses in student writing (National Board for Professional Teaching Standards 2003b). While these last three are "teacherly" activities, the first three are more the kind of exercise a teacher would ask a student to complete.

The purpose of the Next Generation assessment center questions is to determine a teacher's foundational knowledge in the field—although it is not clear that these questions provide any more or better information to the evaluators than the Original Format questions did. The shortened time frame and the removal of advance materials for preparation, however, indicate that the evaluators are more interested in seeing what a teacher can do on the fly than what they can do with thoughtful preparation. Ironically, much of what they ask for in these impromptus

is material that a teacher would likely take more than 30 minutes to prepare.

It is also ironic that while Delandshere and Petrosky advocate more complex assessment to go with complex performances, the National Board makes the tasks less complex. And the scoring procedure has not changed. Each entry, both in the portfolio and from the assessment center, is evaluated separately on a 4–point rubric that defines best practices according to National Board standards. This is the procedure that Moss suggested produces questionable results (1996, 25), and it remains the procedure, despite Delandshere and Petrosky's findings that numerical ratings encouraged evaluators to play a kind of matching game, looking for surface evidence of particular features listed on the rubric, rather than evaluating the performance in depth and as a whole. When Delandshere and Petrosky asked raters to write interpretive summaries, on the other hand, they paid more attention to the performance as a whole, took more notes, and referred to more specifics in their summaries, ultimately providing more useful feedback to the teachers seeking certification (1998, 21). Such practices would be in line with National Board goals and objectives, which focus on reforming education through specific attention to teaching (National Board for Professional Teaching Standards 2003d). Given the research and the fact that Delandshere and Petrosky's procedure was developed specifically for the National Board, the Board's decision to stick with the established, statistically safe route is particularly questionable.

All of these scholars—Linn, Baker, and Dunbar; Moss; and Delandshere and Petrosky—argue that more complex performances require more complex assessment procedures. Complex performances include writing in the manner compositionists teach it—complete with planning, drafting, and revision, time to consider ideas, and the opportunity to gather and respond to feedback from readers. But the messages from these scholars and the alternative practices they advocate have not been adopted by the establishment in K-12 education, where educational measurement theory has the greatest influence. For example, while most states do use direct assessment of writing in their state-level testing, they also still rely on a holistic or criterion-referenced scoring system that requires rater training and calibration, rather than the kind of interpretive or hermeneutic approach these scholars advocate (*Profiles of State Education Systems* 2001). And in these assessments, traditional reliability still functions a limiting condition in assessment; training and calibration of readers to produce reliability results is still a key (and explicit) issue in many of the state descriptions of their assessments.

Unlike Messick's validity work, the work of these theorists, however, is on the cutting edge; it is not the norm, as it appears in the most recent *Educational Measurement* (Linn 1989), the authoritative text for defining and designing assessment procedures, or the most recent *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education 1999), or in current textbooks, such as *Measurement and Assessment in Teaching* (Linn and Gronlund 2000). The current norm in educational measurement is to force-fit traditional conceptions of reliability with validity and its "new" emphasis on construct—usually through some idea of "generalizability," which is an attempt to deal with the increased error that appears as assessment instruments become more complex. This process is awkward at best and becomes nearly impossible in performance assessments such as the direct assessment of writing in anything other than a holistically scored, timed impromptu form. State departments of education know this, and so they keep the holistically scored impromptus, despite evidence that more complex assessments produce better and more useful results.

DEVELOPING THEORIES

Despite its dominance in education and in composition studies, educational measurement theory is not the only framework available for evaluating writing assessment practices, and compositionists have explored alternative theories for some time. Possibly the earliest instance appears in Anne Ruggles Gere's "Written Composition: Toward a Theory of Evaluation" (1980), where she outlines a theory based on "communication intention" and formal semantics. References to this theory, however, are rare, and it seems that Gere's ideas appeared too early to be taken up by compositionists. More recently, Deborah Holdstein (1996) uses feminist theory to challenge notions that holistic scoring and other instruments and procedures can be "value-free." Even White, ultimately a staunch advocate of educational measurement theory, argues that reader-response theory supported of holistic assessment in his 1984 edition of *Teaching and Assessing Writing*—although he largely drops this argument in the 1994 edition, probably because of the waning support for holistic assessment at the time and the pragmatic shift in the book's emphasis. Most of these efforts, however, languish in obscurity, even in writing assessment circles. In the last few years, however, there have been some substantial challenges to reliability specifically, and two, in particular—Richard Haswell's (1998) work in categorization theory, and Bob Broad's (2000; 2003) work in

support of hermeneutics—present significant alternatives to the dominance of psychometric principles and are drawing some attention.

As part of his work on the WSU assessment program, Haswell explores categorization theory, which he explains most fully in "Rubrics, Prototypes, and Exemplars: Categorization Theory and Systems of Writing Placement" (1998). In this essay, Haswell outlines three types of categorization—classical, prototypical, and exemplar—which appear in current literature in psychology, social science, and language analysis. Classical categorization matches a new instance with a category defined by a pre-existing set of features; we identify a book as a novel because we "determine that it is long, fictional, and prose" (1998, 245). Prototypical categorization matches a new instance to an idealized fictional representative of a category; we identify a book as a novel because we think of novels as being "about 300 pages long" with plenty of action and dialogue (1998, 246). Exemplar categorization matches a new instance to a memory of a similar instance; we identify a book as a novel because it looks like a novel we have read recently (1998, 247).

Haswell argues that, although holistic assessment that relies on rubrics professes to rely on classical categorization, raters actually behave as if they were relying on the prototypical variety. While rubrics identify key features and the quality of those features, in actuality, no anchor essays—or actual essays for that matter—"are true to the scale of quality pictured by the rubric" (1998, 242). Essays that may be excellent in some areas are weaker in others, and it is almost impossible to find the kind of consistent rise in levels of quality across features that "true" holistic scoring requires. Since there is no such clear match to characteristics in actual holistic scoring sessions, instead Haswell finds that raters tend to look for indications of similarity with ideals (the prototypical variety) (1998, 247–48). He also argues that if left to their own devices—without norming and rubrics—raters use both prototypical and exemplar types of categorization (1998, 248).

Haswell's use of categorization theory presents a strong challenge to reliability. A key premise behind reliability is that any assessment should be interchangeable with any other assessment, given appropriate training of raters. Categorization theory, on the other hand, suggests that those with relevant experience will be able to make better placement decisions than those with little or no expertise precisely because of their familiarity with the context. In this way, he argues in favor of a kind of limited subjectivity in the process of assessment, which stands in stark contrast to the efforts at objectivity engendered by reliability. The replicability of the

decision across raters is less important than the attention to experience. It is interesting to note, however, that while Haswell provides reliability data in this piece only to challenge it, he does attempt to provide reliability data in *Beyond Outcomes* as part of his justification for the two-tiered method.

Categorization theory, as Haswell outlines it here, provides strong theoretical support for the two-tiered placement procedure at WSU, and Haswell makes no bones about the fact that he turns to this theory specifically to support the practice (1998, 232). But his theory supports a *placement* practice in which there are obvious or majority decisions: most students are placed in the mainstream composition course. It is not so clear how this theory would apply to situations where there is no clear majority decision—grading, for example—or where the instrument of assessment is more complex, as in portfolios. That is, it is not clear that categorization theory could work as an assessment theory that applies beyond limited placement situations, and further research is necessary to explore the scope of its application.

Some of the issues I am raising appear in Bob Broad's development of grounded inquiry theory as a method for developing assessment procedures as well as researching assessment practices. In his work at "City University," Broad outlines a theoretical approach that draws specifically on hermeneutic theory in assessment, which he characterizes as an "emerging paradigm" (2000, 231). Participants in the first-year composition program at CU wanted to standardize their assessments of mid- and end-of-term portfolios. Examining conflicts in their attempts to do so in their norming and communal, "live" trio evaluation sessions, Broad argues that the desire to be both fair and consistent on the one hand and sensitive to complexity and diversity on the other threw their evaluative conflicts into relief. He discusses how one group in particular worked hard to either identify a representative "C" portfolio or develop a list of characteristics of such a portfolio, so that they could then compare each of the actual portfolios. Theoretically, this work would clarify the pass/fail decisions the group needed to make. However, the group was unable to identify a representative portfolio because they could not agree on a selection, and they were unable to develop a clear set of characteristics because they kept turning up anomalous cases in the actual portfolios they were evaluating (Broad 2000, 234–36).

Although Broad does not use Haswell's work on categorization theory to support his research, his work suggests its application, and consequently suggests a difficulty in applying this theory to more complex instruments.

In Broad's descriptions of the work the groups did to reach consensus and standardize their assessments, the groups tried to use classical and exemplar categories to make their decisions: they tried both to develop lists of characteristics and to choose examples of a bare minimum "pass" to guide their assessments. The groups found, however, that neither provided the kind of guidance necessary. On the one hand, this would suggest that the norming efforts were the point of failure, a conclusion that Broad explicitly addresses (2000, 232–34). However, would prototypical categorization have been any better? Fictionalized ideals generate their own problems, particularly when the real objects of assessment seem to all be "weird cases" that defy efforts at standardization (Broad 2000, 232–38). Broad's work suggests that categorization theory would not have helped the assessment process at CU.

Broad's efforts in this piece, however, are focused on challenging the use of norming to achieve standardization. He argues that what appear to be problems or errors from the psychometric approach—the one that provides the theoretical support for norming—are actually advantages and strengths from a hermeneutic perspective. For example, hermeneutics supports the inclusion of dissent in the process of making judgments, an inclusion that psychometrics works hard to eliminate. Broad's use of hermeneutics explicitly draws on contemporary assessment theory's challenge to reliability, a challenge that, as I have argued, has largely been ignored in educational assessment. However, Broad's use of grounded inquiry theory does provide a method for getting at the values that evaluators hold, as well as providing a research methodology, that differs substantially from educational measurement theory, particularly in terms of methods for examining the content of an assessment.

Grounded inquiry provides a method for determining what evaluators value, but Broad relies on hermeneutics as developed in educational assessment to support a method for the actual assessment—a reliance that would recommend Broad for membership in the co-optation group. In this way, like Huot, Broad ends up with one foot in each camp and the implicit argument that composition studies cannot do it alone. In his later work, Broad (2003) focuses much more on the contextually determined development of assessment criteria through a process he calls "Dynamic Criteria Mapping." He uses DCM to pull together hermeneutics and grounded inquiry in a way that elides the traditional distinction between content (validity) and method (reliability or hermeneutics). But Broad does not intend DCM to provide the process for assessment. Instead, it provides a map of the criteria that instructors and/or evaluators actually

use in their assessment of student work. The result is a more honest sense of a program's values, but not necessarily a process for assessment or the criteria for evaluating it.

Both Haswell's and Broad's work have a lot of potential, and both provide interesting alternatives to the dominant theory. While I return to both their practices and analyze them in a different theoretical framework in chapter seven, as they present the theories themselves, I think both sell themselves short. That is, I believe the theoretical models that each relies on provides only a piece of the picture, when a full image would do much more. Haswell overtly challenges reliability, but only in a limited situation. Broad's challenge to reliability stays within educational measurement theory, but more importantly, I feel his "challenge" to validity does not provide criteria for evaluating an assessment, only for making explicit and expounding the values of a writing program. By only taking on parts of the theoretical problem, these approaches, whatever their merits, remain in danger of being reconnected to educational measurement principles, or of having little influence.

The composition studies community has yet to fully develop theoretical principles for writing assessment. Practices described by Haswell and Broad notwithstanding, the majority of current practices reflects an objectivist epistemology and remains more heavily influenced by educational measurement theory than by compositionists' own principled understanding of writing and learning to write. Theories of writing assessment informed by literacy scholarship would provide more relevant sets of standards by which to judge the value of specific assessment practices. Moreover, such theories would enable compositionists to present principled arguments about the demands of writing education and evaluation to governmental bodies, testing agencies, educational institutions, and other entities with an interest in assessment procedures and results. If composition scholars wish to align literacy theory and assessment practices, and if we wish to influence writing assessment practices beyond our individual programs' walls, we need well-developed and grounded theoretical principles for writing assessment built on what we know about literacy.