3
WRESTLING WITH POSITIVISM

Despite the tension between the two, the clash between the objectivist paradigm of assessment and the contextual paradigm of literacy has not simply resulted in an impasse. Large-scale writing assessment exhibits historical and ideological tendencies toward an objectivist epistemology, and while those tendencies are politically weighted, they are neither inescapable nor inevitable. At the post-secondary level researchers, scholars, and administrators have been implementing some alternative approaches to large-scale writing assessment more closely aligned with contemporary theories of literacy. Although these alternative methods are still subject to the influences of objectivism, they are genuine attempts, if not to move outside of this paradigm, then at least to maneuver within it.

Although they appear relatively infrequently as actual writing assessment practices at any level of schooling, two models in particular—the expert reader model and constructivist evaluation—suggest ways of using notions about situated reading and social constructionist principles to inform writing assessment. These models, however, are both promising and limited. While they suggest methods for assessing writing compatible with at least some of the principles of contemporary literacy scholarship, both remained grounded in, or at least beholden to, theoretical principles developed outside of composition studies—at least as they are currently justified in the scholarship. Of the two, the expert reader model has closer ties to composition studies in that this assessment procedure has been developed within the field, but in the literature, it often remains circumscribed by principles of educational measurement. Constructivist evaluation was developed as a social science research methodology to counter objectivist shortcomings and conceits regarding research subjects and results. While this model offers a promising move away from objectivist epistemologies, its research orientation limits its practical and theoretical applications to writing assessment. Both, however, are worth exploring for the ways in which they work counter to the objectivist principles that tend to strangle large-scale assessments.

THE EXPERT READER MODEL

The first contemporary model for large-scale post-secondary writing assessment relies on expert readers—readers with significant prior experience with the assessment decision to be made. This model has grown out of dissatisfaction with the process and effects of "norming" or "calibrating" readers for holistic scoring. Instead of training readers to read "correctly," the expert reader model relies on the experience and expertise of the evaluators to render sufficiently accurate judgments that correlate well with each other. This model breaks with the procedural mainstay of holistic scoring—norming—and this would suggest at the very least a corresponding break with the principle of reliability. However, the discourse surrounding this model indicates ongoing ties with educational measurement theory, particularly in the way that discussion of the use of "experts" implies a kind of pre-assessment norming through experience and knowledge to ensure that their readers will arrive at reliable and valid judgments. That is, rather than engendering alternative principles for assessment, they appear, for the most part, to accept the principles of validity and reliability, particularly the latter, as a sort of ground zero, and to work on alternative practice, rather than alternative theory. The focus, for example, in the two best known expert reader models—the first at the University of Pittsburgh and the second at Washington State—has been not on disputing these principles, but rather on drawing attention to the assessment decision and decision-makers.

In 1993, William L. Smith presented research examining eight years of placement assessments at the University of Pittsburgh. This well-documented and detailed study specifically analyzes the reliability of raters' judgments. Smith compares the inter-rater reliability of raters from within Pitt's composition program and from other university composition programs. He demonstrates that while statistically acceptable reliability can be achieved through training readers—what Smith calls "calibration training"—even better reliability can be achieved by raters who have recently taught the courses in which they are placing students—what he calls "'having taught' training" (1993, 204). That is, those with recent immediate experience with the consequence of the assessment decision are better able to reach agreement about the results of the assessment than those without such direct experience, and thus expert readers are more reliable. Based on the results from his work with expert readers, Smith revised the rating procedure at Pitt so that only teachers who had taught a course most recently could make placement decisions about *that*

course (e.g., a teacher who had taught English A the previous semester, but not English B or C, could make decisions about A, but not about B or C).

Smith focuses on "reliability" as a key concept throughout his study, but he replaces its companion term "validity" with "adequacy" (1993, 144). He explains that the issue of validity in the holistic scoring of placement essays "has not been sufficiently addressed" and that "validity carries a considerable amount of baggage" (1993, 144). By substituting "adequacy," Smith brackets questions about the best or most important kind of validity being debated at the time, and he focuses, instead, on the question of how "adequate" or appropriate the placement decisions are. To judge the adequacy of the placement decision, Smith uses a number of indicators, primarily classroom teacher perception of the decision and final course grades. These indicators, he argues, supply the necessary information about whether or not the placement decision is "correct," and he points out that for placement purposes, "adequate" is a sufficient condition for success.[18]

Like the Pittsburgh study, Washington State's program relies on expert readers. WSU's assessment and writing program is perhaps the best documented, at least in terms of published materials. In the most recent piece, *Beyond Outcomes: Assessment and Instruction Within a University Writing Program* (Haswell 2001c), the contributors, all participants in the writing program at WSU, describe the development of their nationally recognized writing assessment program, which includes coursework, assessment, and writing practice throughout students' college careers. Although used in a couple of places in the assessment program, WSU's use of expert readers is most clearly explained in the discussion of the placement assessment for first-year composition. Because of local constraints, including budgetary limitations and reader turnover, program directors developed a two-tiered placement system in which members of the faculty read the essays first to answer a single question: "did the student obviously belong in regular freshman composition or not?" (Haswell 2001b, 42). The remainder are passed on to more experienced instructors and program administrators who make decisions about all other courses, including basic writing, one-hour labs, ESL courses, and exemptions.

This process is also described in an earlier essay by Richard H. Haswell and Susan Wyche-Smith,[19] "Adventuring into Writing Assessment" (1994), revised and included as chapter two of this collection, which is worth discussing for the language choices the authors make. In this earlier version, Haswell and Wyche-Smith characterize their assessment program as

focused on validity: "In that conflict between reliability and validity which lurks under the surface of all assessment accounts, we would put our money on validity" (1994, 229). They arrive at this point through their dissatisfaction with and dislike of holistic scoring, arguing that holistic scoring, which promotes reliable scores, also mitigates against the careful reading necessary for placement and diagnostic decisions. While their argument tends to conflate holistic scoring and reliability, their point is a good one: it is more important that a test predict students' needs than that scores return a statistically satisfactory reliability coefficient.

When Haswell and Wyche-Smith describe their own follow-up work on the procedure in the 1994 piece, they use psychometric terms: "we spent much of our time investigating the *validity* of our rating system, the effects of actual placements as seen through the eyes of students and their teachers, and the *reliability* of the prompts" (1994, 234, my emphasis). The use of this vocabulary is predictable: "validity" and "reliability" are the key terms of the dominant assessment lexicon. What is more interesting, however, is the subtle shift in the deployment of these terms. "Reliability" is usually applied to rating systems and test results, not specifically or primarily to the prompts for assessments. "Validity," the term of the set more often connected to writing prompts at the time, is here applied to the rating system.

The authors suggest an explanation for this shift earlier in the essay when they claim that instead of focusing on scoring reliability, their assessment emphasizes "'instructional validity,' where a test elicits a response that can be matched with the writer's past and future coursework" (Haswell and Wyche-Smith 1994, 229). This contention—in which one type of validity is used to dismiss reliability—is neither developed nor satisfying as an explanation for using these terms in this manner. It seems likely that in the absence of a more acceptable assessment vocabulary, the authors are trying to stretch the connotations of the existing terms. These two passages, however, are not equal to the task of revising a nearly century-old lexicon, nor do they seem intended to do so. "Adventuring into Writing Assessment," as the title suggests, is primarily a description of practice, not a revision of theory. The result is a model that draws attention to the decision to be made and who is best able to make that decision rather than to the procedural technicalities. The inclusion of "validity" and "reliability" in this case seems more an acknowledgement of accepted principles than an attempt to redefine them.

Even though Haswell and Wyche eliminate most of the psychometric terms from their revision, this language follows the authors in the 2001

collection. They still talk about "training future raters" (2001, 23), and about the need to develop a system that would "maintain reliability and validity from session to session" (2001, 18), but there is little specific reference to psychometric principles in their revision for this chapter. There are, however, references sprinkled throughout the book, the bulk of which appear in the essay by Galen Leonhardy and William Condon (2001). In this piece, the authors examine the "difficult cases" in the junior portfolio piece of the assessment program: transfer students, ESL students, differences between the rater's background and the student's work, etc. Leonhardy and Condon use educational measurement theory to bolster their claims that the WSU assessment program is theoretically sound. They begin with the premise that "[a]ny assessment must meet the basic requirements of validity and reliability", though they also claim that success can only be determined if there is "improvement" (2001, 67). They argue that the WSU program "allows faculty raters to make highly reliable decisions about writing samples that possess a high degree of validity," and they point to the previous two chapters—"The Two Tier Rating System" and "The Obvious Placement," both by Richard Haswell—as demonstrations of this point (2001, 68). With the reliability and validity of the program established, it seems, the authors are free to make their argument for how they have examined and improved upon the difficult cases.

One of the two chapters Leonhardy and Condon point to as supporting their validity claim is the theoretical discussion of the two-tier placement system by Haswell, "The Obvious Placement: The Addition of Theory" (2001a). Here Haswell argues that the theory followed the development of the procedure, a practice that he claims "may be its normal mode" (2001a, 57). In this case, he uses categorization theory—the process of "sorting things into conceptual boxes"—to explicate the procedure (2001a, 57). He distinguishes the kind of categorization in the two-tier program from holistic scoring by pointing out that the latter is "classical categorization" which depends on clear boundaries between categories, while the former is "prototype categorization" which relies on a fuzzier sense of what is "typical" in a category (2001a, 57–9). Haswell develops categorization theory more fully elsewhere (1998), which I, in turn, explore more fully in the next chapter, but his purpose in this section of *Beyond Outcomes* is to explicate a theory supporting WSU's placement system, and here he sidesteps the issues of reliability and validity. He argues that the two-tiered system "finesses the double-pronged threat of cost-efficiency and legitimacy" that has undermines other assessment

systems (2001a, 55); "legitimacy" here implies validity and reliability, particularly given Leonhardy and Condon's comments.

Like Smith's substitution of "adequate" for "valid," Haswell's outline of an alternative theory, as well as his use of "legitimacy," provides a fairly clear indication that the dominant vocabulary for assessment is flawed, or at the very least insufficient. The thorough development of alternative definitions or of an entirely different set of terms, however, is beyond the scope of these projects. In both cases, the focus remains on the procedure. Smith, for example, spends significantly less space explicating "adequacy" than exploring challenges to "reliability" (less than ten pages of sustained discussion out of an essay of more than sixty pages), and his study is cited primarily for the ways in which it champions teachers' expertise and challenges the reliability of holistic scoring. Haswell's theory chapter makes theoretical claims only for the placement decisions made within the two-tier system. Moreover, it is only one of 15 in a book whose purpose is to describe, to provide "an unusually frank and scholarly look at the development, the structure, the problems, and the effectiveness of a robust, university-wide set of writing programs" (Condon 2001, xvii), and the expert reader model is only one element in this set of programs. The primary purpose of the expert reader model is not to circumvent or redefine the lexicon of writing assessment; however, it may suggest such revisions, and in that vein, I will return to WSU's program in chapter seven.

The expert reader model is based on the notion that those most familiar with the decision to be made are best able to make that decision and that they do not need additional "training" to do so; the studies described above support this "common sense" conclusion. "Common sense," however, also suggests that those most familiar with the decision to be made may be the least able to act wisely precisely because of their proximity and investment; this was the claim of much of the early criticism of assessment that is done by teachers. "Common sense," that is, does not provide adequate justification for an assessment model. Educational measurement theory historically has provided such legitimation, and both the contexts described above continue to rely on that theory—even as the practices challenge it. Described in terms of that theory, expert reader assessments seem to have both increased validity and increased reliability because the reliance on knowledge about writing practices promotes informed evaluations that are consistent within the given context.

When viewed within the framework of educational measurement principles, the expert reader model can be understood as exchanging the artificial calibration necessary for inter-rater reliability in holistic scoring

for an implicit alignment between the context for the writing and the context for the assessment, which provides sufficient reliability. While the model itself offers interesting possibilities for alternatives to educational measurement principles, the political weight of the current terminology keeps its proponents justifying this model in psychometric terms. The model's potential is unnecessarily bogged down with principles that are theoretically incompatible.

Legitimation via the terms of educational measurement is a double-edged sword. By calling on "reliability" and "validity" to justify the expert reader model, scholars turn over judgments about the value of writing assessment methods to those outside of composition. Historically, such outside legitimation has encouraged the devaluation of compositionists' expertise, a consequence that contradicts the confidence in and dependence on expertise integral to this model. Moreover, as I discussed in chapter one, this lexicon frames writing assessment in terms appropriate for an objectivist paradigm. Individual programs may be able to function outside the established norms, but they do so without strong theoretical corroboration. The expert reader model offers a promising alternative to the practice of norming, but its continued development would seem to require theoretical principles more in keeping with those of research and scholarship in the fields of composition and literacy studies.

CONSTRUCTIVIST EVALUATION

Constructivist evaluation, the other cutting-edge assessment model, relies generally on social constructionist principles but derives most specifically from the work of Egon G. Guba and Yvonna S. Lincoln in *Fourth Generation Evaluation* (1989). Constructivist evaluation contends that any evaluation should be a thoroughly collaborative and contextualized project and that all who are affected by an assessment should have a voice in the process. The idea that assessment—perhaps the most top-down, authoritarian aspect of composition pedagogy—could be socially constructed makes this model particularly appealing to composition scholars who tend to accept social constructionist principles in both their theoretical work and in their day-to-day pedagogical practices. However, the practical and political limitations to this model call into question its value for large-scale writing assessment The logistical feat alone of gathering input from all those with an investment in any given assessment is daunting. But while practical problems can be overcome, compositionists should question whether or not all those affected by an assessment should decide how that assessment should proceed. The inclusion of students' concerns in the

evaluation procedure sounds attractive, but students are hardly the only other participants. Constructivist evaluation, that is, raises both practical and political issues that call its efficacy into question. Its theoretical similarity to generally accepted principles in composition studies, however, makes it worth considering as an alternative to objectivist assessment.

Guba and Lincoln present constructivist evaluation explicitly as a countermeasure to objectivist research and testing. They point out that evaluation has relied almost entirely on a scientific paradigm, "grounded *ontologically* in the positivist assumption that there exists an objective reality driven by immutable natural laws," i.e., some fixed reality, "and *epistemologically* in the counterpart assumption of a duality between observer and observed that makes it possible for the observer to stand *outside* the arena of the observed, neither influencing it nor being influenced by it" (1989, 12).[20] By contrast, fourth generation evaluation is grounded in "the constructivist paradigm (also called, with different shades of meaning, the *interpretive* or the *hermeneutic* paradigm and, sometimes—erroneously, we believe—the *qualitative* paradigm)" which relies on relativist ontology and a subjective epistemology (13).[21] Guba and Lincoln argue that the assumptions of the constructivist paradigm are "virtually polar" to those of the scientific paradigm:

> For *ontologically*, it denies the existence of an objective reality, asserting instead that realities are social constructions of the mind. . . . *Epistemologically*, the constructivist paradigm denies the possibility of subject-object dualism, suggesting instead that the findings of a study exist precisely because there is an *interaction* between observer and observed that literally creates what emerges from that inquiry. *Methodologically* . . . the naturalistic paradigm rejects the controlling, manipulative (experimental) approach that characterizes science and substitutes for it a *hermeneutic/dialectic process.* (1989, 43–44)

The social emphasis of this model makes it particularly appealing to compositionists who, in general, would tend to accept the idea that knowledge—and by extension, writing—develops in social interaction, and that consequently evaluation is—or should be—a social act.

According to Guba and Lincoln, the society involved in any given assessment is made up of "stakeholders," a term which refers to those who initiate an evaluation, who participate in it and who are affected by it, whether that effect is positive, negative, or in between. Fourth generation evaluation is predicated on the notion that all stakeholders should have a say in constructing any evaluation that concerns them. Guba and Lincoln identify three classes of stakeholders in any evaluation: "agents,"

"beneficiaries," and "victims." Agents are those who produce, administer and use the evaluation; beneficiaries are those who profit in some fashion from the evaluation; and victims are those who are harmed in some fashion by the evaluation. If constructivist evaluation is to be "responsive," it must attempt to address the concerns, claims, and issues of all of these groups (1989, 40–41).

Once a need for evaluation has arisen, fourth generation evaluation proceeds through a process of identification and negotiation. The person or group initiating the evaluation first identifies all the stakeholders and elicits from them their claims, concerns and issues—that is, their construction of the evaluation. The initiator then sets up the negotiation by identifying points of consensus, by providing a context, methodology and agenda for the negotiation of points in dispute, by facilitating the actual negotiation, and finally by generating one or more reports that convey points of consensus, dissensus, and resolution. Those points still unresolved are then subject to further negotiation. Theoretically, fourth generation evaluations never end; they merely "pause until a further need and opportunity arise" (1989, 74).

Guba and Lincoln argue that the limitations of conventional evaluation are replicated in the criteria—such as "validity" and "reliability"—used to determine the worth or "goodness" of the assessment. After describing a set of criteria which would parallel the criteria of "rigor" applied to conventional assessment,[22] they contend that their revised criteria are insufficient for determining the quality of a fourth generation evaluation precisely because they parallel the criteria of conventional evaluation and are thus limited by its positivist assumptions (1989, 74, 245). Moreover, they argue, these criteria—whether the conventional or revised set—are unacceptable because they are primarily methodological criteria and thus serve predominately as an internal check on the process of coming to conclusions. Such criteria do not, for example, provide principles for determining the value of the purpose(s) for or the outcome(s) of any given assessment. Guba and Lincoln point out that "[i]n the positivist paradigm, method has primacy," but that in a constructivist paradigm, method is only one issue among many (1989, 74, 245). Thus, any criteria modeled explicitly on conventional criteria would be inadequate.

In place of parallel methodological criteria, Guba and Lincoln offer a set of criteria they pull together under the heading of "authenticity." They argue that these criteria—"fairness" and ontological, educative, catalytic and tactical authenticities—arise from the assumptions of the constructivist paradigm itself (1989, 245–50). Fairness focuses on the solicitation and

honoring of all the stakeholders' constructions and requires negotiations that level power dynamics. Ontological authenticity refers to the extent to which the stakeholders' own constructions become more informed and sophisticated as a result of the evaluation process. Educative authenticity indicates the extent to which the stakeholders come to better understand the constructions of others. Catalytic authenticity signifies "the extent to which action is stimulated and facilitated by the evaluation process" (1989, 74, 249). Finally, tactical authenticity refers to the extent to which stakeholders are empowered to act as a result of the evaluation process. Unlike conventional criteria, each of these draws attention to the relationships among the stakeholders and changes in the stakeholders' understandings and/or abilities. Moreover, the application of these criteria continues throughout the evaluation as part of the case study record that results from a fourth generation evaluation. Thus, evaluation of the evaluation is an ongoing part of the assessment process.

Presumably, once a negotiation is complete—or paused—the evaluation would have addressed the concerns of all participants, even if they all could not be resolved. A "constructivist evaluation," supported by the concept of "stakeholders," ostensibly levels the playing field, and all participants have an equal say. Guba and Lincoln argue that conventional evaluation "effectively reserves power and decision-making authority" to the clients who request the evaluations and as such, are "not only morally and ethically wrong but also politically naive and conceptually narrow" (1989, 15). Their method aims specifically at redressing the wrongs that inhere to conventional practices.

Sandra Murphy and Barbara Grant suggest ways to enact this constructivist model in writing assessment situations in "Portfolio Approaches to Assessment" (1996). Murphy and Grant describe the conditional nature of constructivist assessment and the effort at methodological consistency it requires. They point out that there is nothing inherently constructivist about portfolios, that portfolios may be implemented in positivist ways toward positivist ends. For example, they argue that standardizing the contents of student portfolios—including requirements for certain genres or demonstrations of specific abilities—reinscribes positivist ideals about objectivity and reproducibility that require stripping the context for writing from assessment practices. According to Murphy and Grant, constructivist portfolio assessment, at the very least, would have to develop directly from the pedagogical context in which the materials for portfolios are generated. Nor does such a contextually aware assessment guarantee a constructivist model: only if the pedagogical practices in the

classroom reflect constructivist values could constructivist assessment follow. According to the authors, for example, the pedagogical context would have to allow students to develop their own assignments—in conjunction with their teachers—and then allow them free reign in choosing the contents of their own portfolios. Furthermore, a constructivist assessment would need to be collaborative, which for Murphy and Grant means that both teachers and students, as stakeholders, participate directly in developing the criteria for assessment and in the process of assessing itself. Although they are not clear about the connection between top-down (i.e., non-collaborative) assessments and positivism, they imply that assessments developed by administrators and imposed on faculty and students reflect an unwarranted positivist faith in the superior, objective knowledge of the expert/manager who stands outside the pedagogical context.

Murphy and Grant are only able to point out a few portfolio projects that embody the principles of constructivist evaluation, including the placement portfolios at Miami University of Ohio. Wholesale imports of fourth generation evaluation, however, are virtually non-existent in practice and rarely even appear in the literature on writing assessment to-date. The greatest obstacles are practical. The process of continual consultation and negotiation is too unwieldy, time-consuming, and expensive for most post-secondary composition programs to manage, a situation that even those advocating Guba and Lincoln's methodology acknowledge. More often, the application of fourth generation principles is partial.

Composition scholars usually limit their use of Guba's and Lincoln's methodology to clarifying and addressing the concerns of the "stakeholders" involved in any given assessment—though the origins of the term often go unacknowledged. For example, in "Power and Agenda Setting in Writing Assessment," Edward M. White presents the primary "assumptions, perspectives, and demands" of the four dominant stakeholders in large-scale writing assessment: "teachers; researchers and theorists; testing firms and governing bodies; and students, especially those from minorities and other marginalized groups" (1996a, 11).[23] He argues that these "stakeholders stand at wholly different positions and are bound to see writing assessment from where they stand" (1996a, 23). His point is that each group needs to understand the positions of the others and at least honestly attempt the negotiation of their differences.[24] It would be more accurate, however, to say that White believes that compositionists need to understand and even concede to the viewpoints of measurement and testing specialists. His audience is, and has been, almost exclusively

composition studies professionals, so when he argues for negotiation, he is telling us—compositionists—to listen to them—measurement specialists.

White's implicit argument that compositionists should pay particular attention to measurement specialists stems from his understanding of a power differential between those specialists and the community of writing teachers. White has argued repeatedly that if compositionists do not take the position of measurement experts and agencies seriously, this outside group will define writing assessment for us. He clearly understands the measurement community to wield greater societal power—enough that they could, if they wanted to, take control of writing assessment. This seems a logical—if somewhat dramatic—conclusion to draw, considering how much influence testing organizations such as ETS have in national affairs. Were ETS to come to the negotiation table as an equal among stakeholders, as they would have to in the ideal situation that Guba and Lincoln describe, they would be relinquishing far more power than any other constituent, except perhaps governmental bodies—which, until very recently, have seemed to be less interested in specific testing practices than in ensuring that assessment occurs. Consequently, there would have to be a significant incentive for members of testing agencies to even sit down. Guba and Lincoln's configuration presumes that all parties participate voluntarily and that they voluntarily check their clout at the door.

There is little reason to believe that testing agencies would do so. For example, while some members of ETS and other major testing agencies do join discussions within the composition community, their numbers are few, and the organization as a whole does not seem to feel any pressing need to address wholeheartedly the concerns of compositionists. Roberta Camp, one of the few, may be the most frequent member of the testing community to join writing community discussions. She and Hunter Breland, both of ETS at the time, contributed essays to the collection *Assessment of Writing* (1996), and Camp also has an essay in *Validating Holistic Scoring* (1993). Although sympathetic to the concerns of writing teachers, she tends to advocate refining the principles of educational measurement with insights from the writing community rather than seriously reconsidering the principles themselves. Moreover, within ETS, at the time she was writing these pieces, she was not responsible for test development or administration (Hamp-Lyons 1995, 449).[25] On the other hand, Breland, a senior research scientist, argues that English teachers should consider using computers for scoring at least the more standardized

aspects of composition, a position that ignores one of composition's most basic tenets: that all aspects of writing, including grammar, are complex and contextually-dependent.[26] Even *Assessing Writing*, a journal whose audience is "educators, administrators, researchers, and . . . all writing assessment professionals," seems to attract little attention from members of assessment organizations.[27] In the eight volumes published to-date, only one essay has been contributed by members of the most influential testing agencies (Wolfe et al. 1996): of the four co-authors, one is from ETS, two are from ACT and one is from Iowa State University. Considering that the original editors of *Assessing Writing* actively encouraged submissions from "a range of scholars and practitioners in the fields of education, composition, literacy, nonacademic workplaces, electronic applications, measurement and administration" (Huot 1994b, 7), the absence of voices from these agencies does not offer much hope that these professionals would negotiate voluntarily from a position of limited power.[28]

Even in light of this obstacle, compositionists have cause to advocate assessment practices that call all stakeholders to the table. A significant portion of the composition community, particularly in post-secondary educational institutions, actively adopts a social constructionist philosophy in both pedagogy and research, so at least in theory, constructivist evaluation would match our beliefs about the ways in which meaning is made. Student-centered and collaborative pedagogies advocating the central position of students in the classroom would likewise intersect well with constructivist assessment, which would necessarily include student voices. Moreover, from our standpoint, it would certainly not hurt to have a forum in which measurement specialists and governmental agents would be required to pay attention to the positions of teachers and composition scholars.

Yet even if such leveling could occur—and I would argue that the practical problems alone are formidable—I am not at all convinced that post-secondary compositionists should want a level playing field. Composition has historically occupied the role of remediator not only in English Studies, but also throughout higher education.[29] The idea that composition is remedial contributes to the misapprehension that writing instruction—and, by extension, writing assessment—requires no special expertise. Compositionists have been asserting their disciplinary expertise for some time and have made some headway, if only indicated by the rising number of both undergraduate and graduate rhetoric and composition programs. Constructivist evaluation—both in its "pure" form as described by Guba and Lincoln and in its simplified "stakeholder"

form—implicitly limits expertise: all stakeholders have an equally legitimate say in all aspects of the assessment. Moreover, because of the status of composition in most post-secondary institutions, the concept of "stakeholder" is subject to inflation. As it is, composition is probably the most frequently mandated core requirement. The mandate generally comes not necessarily from composition programs or even English departments, but often from programs and departments throughout the university. That is, bringing all stakeholders to the table suddenly means that faculty and administrators from any and all departments can and should have a seat at the table—an evaluator would be remiss if s/he did not make a place for all of them—and the weight of compositionists' expertise is even further diminished.

A level playing field carries a price that, I would argue, composition should not pay. This is not to say that we should reject the concerns of other interested parties, but rather that their positions should inform ours, with ours occupying a central position rather than one among many. The members of any other discipline—including mathematics, which houses the other predominant "core" requirement—are able to define their own standards and values, and expertise in any field other than composition tends to carry with it the presumption of such evaluative knowledge. If educators and disciplinary professionals are doing a good job, those standards and values are the product of an ongoing community discussion that incorporates the concerns of interested parties. But the heart of the community consists of those with expertise. Constructivist evaluation potentially negates this expertise, and for composition the effect is magnified. This might explain part of why the composition community has not warmly embraced constructivist evaluation, in spite of its attractive counterstatement to positivism.

THEORETICAL NEED

Whatever their shortcomings, both the constructivist model for evaluation and the expert reader model—which embraces expertise in a manner antithetical to constructivist evaluation—provide examples of assessment practices more or less in tune with composition theory and pedagogy. In the field of writing assessment, such models are relatively rare, though they are becoming increasingly prominent in terms of the research they generate, if not prolific in terms of actual practice. Although promising as challenges to conventional practice, however, neither of these models has more than one leg to stand on. Constructivist evaluation, when applied to composition studies and writing assessment, ignores very real

power dynamics and practical constraints. The expert reader model, as it is currently presented, relies on educational measurement principles for justification.

Still, these models have generated additional research and attention, and the limits of practice notwithstanding, they have a lot to offer. But, I would argue, the contributions will be limited without additional theoretical work. While constructivist assessment provides a theoretical model, it does not adequately meet the purposes of writing assessment: Guba and Lincoln are more focused on research, and while assessment can be understood as a kind of research, its needs are more particular than the application of this model allows. The expert reader model comes across first and foremost as a practice; the theoretical justification comes after. The use of conventional educational measurement theory to justify this model produces unsatisfactory results, and the use of categorization theory, which I will say more about in the next chapter, is not clearly applicable beyond the specifics of the expert reader model. Moreover, categorization theory is unlikely to be applicable to all large-scale assessment situations, and it may not translate into generalizable principles. The emphasis on practice over theory—as much as the two can be separated—either maintains the status quo where educational measurement remains in power, or leads compositionists toward a situation where each practice is supported by its own theory. Neither of these is conducive to effective and long-standing change.

Arguably, we are looking at a paradigm shift. These alternative practices, especially the longevity of and ongoing research into the expert reader model, suggest that the current dominant paradigm of educational measurement theory—as we understand and apply it—cannot answer the questions we currently have about writing assessment. We want, for example, to understand how assessments can affect curriculum and pedagogy positively. We want to develop assessment practices that allow for actual reading practices. We want to find ways to be fair to students from diverse backgrounds that are still sensitive to our course and program objectives. For some time now, educational measurement theory has been pushed to address these concerns, but I believe that it is not up to the task. This tension—between the needs of compositionists in writing assessment and the demands of educational assessment theory—shows up in the contemporary practices. It is all the more apparent when we look at the theoretical work both in this field and in educational measurement, which is the work of the next chapter.