NOTES

INTRODUCTION

1. The term "scene" better connotes a sense of (discursive) action and struggle than does "surface," which implies a certain level of serenity. The term "surface" also implies that there is something hidden, some depth, a connotation that Foucault's work would seem to reject. For these reasons, I privilege "scene," the term from "Nietzsche, Genealogy, History" (1977) over "surface," the term from *The Archaeology of Knowledge* (1972).

CHAPTER ONE

2. Prior to the Boston report, writing instruction had been treated as a literary endeavor, involving reading and imitating "masters" of literature. The Boston tests shifted this literary emphasis toward an expository one in the way they used writing to describe and analyze. Literacy standards, however, did not change at the same time, and the Boston report ignores the fact that the writing most valued at the time was literary or creative, and not the kind asked for in the written tests (Witte, Trachsel, and Walters 1986, 17–18).
3. The report, authored by the examiners who implemented this experiment in written examination, provides data about the number of students assessed and the types of questions asked; Horace Mann, the editor of the *Journal* comments more broadly on the scholarly and practical benefits of these exams. Since the section that concerns contemporary writing scholars appears to be authored by Mann, the article is regularly attributed to him—although different parts appear to be written by different authors. In crediting Mann, I am following suit.
4. Frederick James Kelly (1914) summarizes a number of studies that demonstrate the range of variation in teachers' comments and provides his own study which concludes that composition scales used in conjunction with training narrow the range considerably. C.C. Certain (1921) uses studies such as these to argue that teachers should forego all elements in tests that require subjective judgments about value. A.A. Roback (1921) takes an alternative position, arguing that while the variability is well documented, it is also inevitable and not entirely undesirable.

5. See, e.g., Ballou (1914); Breed and Frostic (1917); Certain (1921); Dolch (1922); Hudelson (1916); Johnson (1913); Kelly (1914); Trabue (1917); and Willing (1918).
6. Trabue (1917) circumvents this problem by remaining silent on the subject of what his scale is supposed to measure. Ballou (1914) provides fairly lengthy summaries of raters' comments on each essay in the scale in an attempt to describe "merits" and "defects," but he provides no clear definition of these terms.
7. A search through the *Education Index* from the 1930s on reveals a dearth of research in direct assessment until approximately the mid-1960s when activity picks up again.
8. See Mary Trachsel's *Institutionalizing Literacy: The Historical Role of College Entrance Examinations in English* (1992) for a detailed account of the history of the College Board and its examinations.
9. There are multiple kinds of validity evidence, some of which I discuss later in this chapter and in chapter four. The distinctions among them are not relevant here, but some good sources exist that explain the differences. For a basic overview of testing theory, see Howard B. Lyman (1991). For a more thorough introductory discussion, see the textbooks by Anne Anastasi (1982) and Robert L. Linn and Norman E. Gronlund (2000). For specific discussion about the issues surrounding the concept of validity, see the essays in Howard Wainer and Henry I. Braun's collection *Test Validity* (1988) and Samuel Messick's "Validity" (1989b).
10. White presents his position in a number of other articles and books chapters, including "Holisticism" (1984), "Pitfalls in the Testing of Writing" (1986), "An Apologia for the Timed Impromptu Essay Test" (1995), and Chapter 13 of *Teaching and Assessing Writing* (1994c), entitled "The Politics of Assessment: Past and Future." I have used "Holistic Scoring" (1993) because I find that he makes his case most clearly, fully, and elegantly here.
11. There are also practical problems: "Essay tests can be read holistically . . . at an average rate of 25 or more an hour (for 45–minute essays). But portfolios can, and often do demand, an hour or more apiece" (White 1993, 101–2). The problem is cost. If nothing else, portfolios cost more, and in ages of downsizing and budget cutting, the bottom line directly influences choices about procedures.
12. The essays by Pat Belanoff (1994), White (1994b), Elbow (1994), James A. Berlin (1994), and Brian Huot (1994a) in the *New Directions* (Black et al. 1994b) collection are notable exceptions, as are the Foreword by Elbow in Belanoff and Dickson's *Portfolios* (1991) and the essay by Huot and Williamson (1997) in *Situating Portfolios* (Yancey and Weiser 1997).

However, these essays represent a small fraction of the contents of each of these collections.

13. Guba and Lincoln's text was published in 1989. There is only one reference to it in Williamson and Huot (1993) and only eight in the first eight volumes of the journal *Assessing Writing*. There are a few references to it in Yancey and Weiser (1997), Haswell (2001c), Huot (2002), and Broad (2003). There are no references at all to it in the essays in *New Directions in Portfolio Assessment* (Black et al. 1994b) or in the second edition of *Teaching Writing* (White 1994c). Yet Guba and Lincoln's text is probably the best known text from outside composition on alternative assessment paradigms.

CHAPTER TWO

14. Given the parallels, it is interesting to note that Trachsel does not cite Street even though, his work was originally published in 1984.
15. De Castell and Luke limit their study to North American literacy and deal with both the United States and Canada in this article, arguing that literacy instruction in both followed largely parallel lines. They point out the distinctions between these cultures as necessary, and where they have, I have relied only on their arguments regarding literacy instruction in the United States since that is the context for my study.
16. Ironically, the reference argues that "[s]tatistics show that literates contribute a larger percentage of their class to the criminal ranks than do the illiterates."
17. See Luke (1988) for a thorough discussion of the development of basal readers.

CHAPTER THREE

18. It is ironic that this research appears in Williamson and Huot's collection, entitled *Validating Holistic Scoring for Writing Assessment* (1993).
19. In *Beyond Outcomes,* Susan Wyche-Smith appears as Susan Wyche.
20. Guba and Lincoln rely heavily on italics throughout their text. All italics are theirs unless otherwise noted.
21. They also occasionally use "naturalist," a term from their earlier work which they dropped in favor "constructivist" because of both the unwanted connotations of the former and the greater appropriateness of the latter. From time to time, "naturalist" or some derivative appears in the more recent texts, usually when the ideas and/or the text are pulled directly from the earlier work.
22. Specifically, they offer "trustworthiness" as a category of criteria parallel to the criteria of "rigor" under positivism, which includes internal

validity, external validity, reliability, and objectivity. In place of internal validity, they offer "credibility" which becomes a level of correlation between the stakeholders' constructed realities and the reconstruction of those realities as represented by the evaluator to the stakeholders (235–36). Such a correlation, they argue, can be obtained through techniques—such as "prolonged engagement," "persistent observation," "peer debriefing," "negative case analysis," "progressive subjectivity" and "member checks" (237–39). In place of external validity, Guba and Lincoln offer "transferability" which correlates "the degree of similarity between sending and receiving [evaluation] contexts," and the level of similarity is the responsibility of the stakeholders wishing to apply the study to some other context (241–42). In place of reliability, they offer "dependability," which requires that the process, complete with changes, be thoroughly documented so that shifts in constructions can be tracked and examined by outside reviewers. Finally, in place of objectivity, they offer "confirmability," which tracks data to its source and which traces "the logic used to assemble the interpretations into structurally coherent and corroborating wholes" through the case study narrative that results from a fourth generation evaluation (242–43).

23. This essay is a slight revision of "Issues and Problems in Writing Assessment," White's essay in the first issue of *Assessing Writing* (1994a). I use the one from *Assessment of Writing* (1996)—the collection of essays he co-edited—because in the anthology, the journal article is acknowledged as the earlier version. There is, however, no substantive difference between the two. Although Guba and Lincoln are cited in both versions, they are cited as an additional reference for challenges to positivism within assessment scholarship and not credited with the term "stakeholders."

24. This is a variation on a theme for White. He has been arguing—as either a primary or a secondary point—that writing teachers need to understand the positions of administrators and testing agencies since at least 1986 with his essay "Pitfalls in the Testing of Writing" in *Writing Assessment: Issues and Strategies.* He has argued explicitly that if writing teachers do not develop assessment procedures which satisfy testing experts and agencies, those entities will take over writing assessment entirely (see, e.g., 1996c; 1994b; 1994c).

25. Camp has since retired from ETS; as of this writing Breland is still employed there as a senior research scientist.

26. See White's "Response" (1996b) in the same volume for more on the response of the composition community.

27. The audience is listed on the front cover of each issue.

28. The current editor, Liz Hamp-Lyons, took over in volume 8 and is actively encouraging a more international perspective, as the adjusted title suggests (Hamp-Lyons 2002). Her focus may be even less likely to draw in educational testing agency scholars whose work focuses primarily on testing U.S. students.
29. See, e.g., Miller (1991) and Slevin (1991) for discussions about composition as remediation and the effects of that designation on post-secondary composition professionals.

CHAPTER FOUR

30. Huot's "Toward a New Theory of Writing Assessment" (1996b) is arguably the first overtly theoretical text on writing assessment to come out since Anne Ruggles Gere published "Written Composition: Toward a Theory of Evaluation" in 1980. Gere argues that "[e]ffective evaluation requires scrutiny of our concept of meaning; anything less will merely tinker with externals" (1980, 58) and outlines a theory based on "communication intention" and formal semantics. It is difficult to find any reference to it in contemporary assessment literature; it is, apparently, a case of premature theorization.
31. In the text of the policy statement, the assumptions I am quoting here appear in italics, which distinguishes the assumption from the explanation which follows it. Since I am only quoting the assumptions, I do not use the italics. I have also omitted the explanations, which are not germane here.
32. In "Power and Agenda Setting in Writing Assessment," White's list, in order, is writing teachers, researchers and theorists, testing firms and governmental bodies, and students, especially those marginalized.
33. For an overview of contemporary validity theory, see Moss (1994) and Messick (1989a). For a full, technical discussion, see Messick (1989b). I discuss these texts in more detail later in this chapter.
34. The italics appear in Messick's article; Huot removes them in his chapter.

CHAPTER FIVE

35. The *Standards* includes vignettes that provide examples of this instructional emphasis, including one class where students choose texts based on their individual interests from among more than 100 young adult novels, and another in which the students—who come from 18 different countries—interview family members about their immigrant experiences and publish their resulting stories in a collaborative portfolio.
36. Originally, the project was financed by the Department of Education's (DOE) Fund for the Improvement and Reform of Schools and

Teaching as part of a nationwide effort to articulate content area standards for elementary- and secondary-level education. This funding was withdrawn after eighteen months, however, based on the finding that "there has not been substantial progress toward meeting the objectives in any of the approved applications" (Diegmueller 1994a, 1). The DOE decided against soliciting proposals for the same project from any other organizations in the wake of a letter-writing campaign prompted by the NCTE and IRA, which argued that these two literacy organizations were the most appropriate to draft the standards, and the NCTE and IRA subsequently decided to fund the project on their own (Diegmueller 1994b, 9). Two years after that decision, the *Standards* was published, along with a series entitled *Standards in Practice*, which illustrates classroom and curricular activities designed to meet high achievement standards, broken down by grade level. The history of the development of the *Standards* is traced in a number of journal articles and in Appendix B of the *Standards* itself. See, e.g., "IRA/NCTE Standards Project Nears Completion" and Diegmueller 1994a.

The DOE sponsored content standards in multiple subject areas, but no other group had its funding pulled, not even the National Center for History in Schools at UCLA, which had originally developed content standards loudly denounced nationwide as too politically correct. Although the history standards received financial support for revisions, the language arts project lost funding, and while there are likely multiple reasons, one in particular stands out. The history standards, in spite of their PC content, still looked like standards, as the critics understood that term. The original history standards—released in 1994—followed the same pattern as those that had come before: laying out specific requirements for content information attached to particular grade levels. The language arts standards did not.

37. Secretary of Education Richard W. Riley and then-President William Clinton were a bit kinder, or at least more politic, than some of the critics. Riley noted, just prior to their release, that the standards had "run into difficulties" (1996), and Clinton pointed out afterwards that attempts to develop standards in both history and language arts were "less than successful" (1996). Before he was justifiably drawn away by the events of September 11, 2001, President George W. Bush looked as if he were going to raise the issue again. At this writing, whether he does or not remains to be seen.

38. The *Standards* has also been subject to criticism from a radical standpoint that challenges the movement for national standards altogether (Shannon 1996). This criticism, however, constitutes only a very small

portion of the negative commentary on the *Standards* and does not reflect the opinion of the more influential respondents.

CHAPTER SIX

39. This definition of "meaningfulness" is far more complex than the one used by Linn, Baker, and Dunbar (1991), which focuses only on the ways in which the assessment is meaningful for students in ways that motivate them.
40. Anthologies even from the early and mid-1990s, for example, abound with unexplicated references to the social construction of knowledge and understanding and to the related concepts of context and community. See, e.g., Bloom, Daiker and White (1996); Bullock and Trimbur (1991); Clifford and Schilb (1994).
41. See Crowley (1998) for a discussion of the continued practical emphasis on current-traditional rhetoric in pedagogical situations, a situation which, it seems to me, has changed some, but not entirely, in the years since.
42. The University of Wisconsin-Milwaukee has since replaced the impromptu examination described by Buley-Meissner and Perkins with a first-year composition exit portfolio system, but the consequences for failure remain as of this writing. Students get three chances each to pass English 101 and 102 portfolio assessment; if they do not, they can not graduate.
43. As a rule, faculty are less likely to be materially harmed as a result of any particular assessment of students—they are not apt, for example, to lose their jobs as a direct result of any single assessment of student writing—and in fact, benefits such as the development of community or additional pay are more probable than any specific damage. I am not, however, arguing that faculty suffer no negative effects as a result of large-scale writing assessment, only that the negative consequences for the student are more immediate and more directly attached to the results of a single specific assessment than the negative consequences for faculty.
44. This is not to say that the purpose of testing in general is an untroubled concept. F. Allan Hanson (1993), for example, points out the American "addiction to testing" which makes our nation's people (not just children) one of the most tested in the world (1993, 1). He argues that the knowledge generated by these tests controls the behavior of individuals by making them complicitous in their own domination—in order to succeed, they must "strive to comply with expectations embedded in tests" (1993, 5). In addition knowledge derived from testing characterizes people "in terms of their achievements and talents, their physical and

mental characteristics, their normalities and abnormalities as measured along innumerable dimensions . . . " (1993, 5). These purposes—control and definition—hardly qualify as admirable, and Hanson does not intend them to be. I return to notions of control and domination later in this chapter and in my conclusion.

45. The criteria for evaluation—whether formal or not—should be developed by those working explicitly within the context so that the continuity among the assessment and instruction is maintained, a point made in many recent assessment-oriented texts, including the NCTE and IRA *Standards*, the CCCC's "Position Statement," Huot's *(Re)Articulating Writing Assessment* and Broad's *What We Really Value*.

46. In chapter four, I discuss Huot's persuasive challenge to the equation of reliability and fairness that White relies upon. Briefly, Huot points out that reliability only indicates consistency among raters—one portion of "fairness" but certainly not its equivalent. He argues that "fairness" must include information regarding "the nature of the judgment itself"; however, he is not clear about what else might be necessary for a "fair" assessment (2002, 88). Huot's project in this section of his article, though, is not to explicate "fairness" but rather to challenge reliability as a theoretical principle appropriate to writing assessment. See also Huot and Williamson (1997).

47. Others who explicitly consider fairness a significant issue include Peter Elbow (1993, 189) and David W. Smit (1994).

48. I would argue that, in practice, determining the purpose and object of assessment would occur roughly at the same time as discussions of criteria and procedures, but for the purposes of explicating the terms "meaningful" and "ethical," I am suggesting they occur sequentially. This is intended to reinforce the idea that determinations of purpose and object should precede determinations of procedure, even if the discussion occurs simultaneously.

49. The term "validity" appears throughout the texts on communicative ethics, but its meaning is less technically oriented than in those coming out of psychometrics. Validity here is used in the sense of legitimacy, authority, or soundness. On the one hand, the usage is troubling in that my work is arguing for a different vocabulary. On the other hand, the link between ethics and validity is somewhat gratifying in its implicit claim that decisions about what is valid are also ethical decisions and not merely rational. "Validity," however, does not appear to carry the same baggage for critical theory as it does for writing assessment.

50. The essay entitled "Discourse Ethics: Notes on a Program of Philosophical Justification" also appears in Benhabib and Dallmayr (1990) published

the same year. The latter version, however, is excerpted and so I am using the essay as it appears in Habermas's *Moral Consciousness and Communicative Action* (1990a). There are some minor translational differences in passages I cite, but none are substantive.

51. Communicative ethics structures argumentation on the notion of intersubjectivity: not only am I a subject, but the other is also a subject. For a discussion of the psychoanalytic principle of intersubjectivity, see Benjamin (1988), especially 25–42.
52. Guba and Lincoln do not cite Habermas or any other theorist working specifically with communicative ethics; they tend to rely on texts that discuss methodological issues and the history and philosophy of science.
53. Benhabib calls the set of rules "the symmetry condition" and the set of relations "the reciprocity condition" (1986, 285). For the symmetry condition to be met, "each participant must have an equal chance to initiate and to continue communication," and "each must have an equal chance to make assertions, recommendations, and explanations, and to challenge justifications" (1986, 285). For the reciprocity condition to be met, "all must have equal chances as actors to express their wishes, feelings and intentions," and all "must act *as if* in contexts of action there is an equal distribution of chances" to act with authority, to refuse authority, to be accountable and to hold others accountable (1986, 285). The symmetry condition refers to the possibility and condition of speech acts. The reciprocity condition refers to the possibility and condition of action.
54. See, e.g., Black et al. (1994a), Sommers (1989), and Yancey (1996).
55. I use the qualifier "published" here because my suspicion is that the majority of *written* work in the area of writing assessment actually deals with issues of meaningfulness; however, since this work appears in the form of unpublished site-specific criteria and scoring guides, it has not been systematically investigated.

CHAPTER SEVEN

56. Absent are specific references to legislatures, university government, or employers, groups that have been named as stakeholders in other scholarship, but these missing stakeholders seem the least directly connected to the purpose of *Beyond Outcomes.*
57. I originally presented the results of this research at the 2001 Conference on College Composition and Communication; my last name at the time was LaCoste (LaCoste 2001).
58. For a more detailed description of the theory of decision logics, see Macoubrie (2003).

59. As both a teacher-participant and a researcher in this project, I have had considerable difficulty with the pronouns in this section. While the use of "we" often feels more comfortable, there is a certain remove from the research now that encourages me to use "they." To clarify the result of my discomfort, I avoid the use of "we" or "I" here unless I am specifically referring to myself.
60. This beginning point is clearer in Broad's 2000 study.

CONCLUSION

61. Interestingly, both do so in footnotes. In "Will the Virtues of Portfolios Blind Us to Their Potential Dangers?" Elbow undermines his utopian vision of minimalist holistic scoring—which would separate the exceptionally strong and exceptionally weak papers from the (rather large) middle ground of unclassified and unscored papers—by pointing out that practical considerations force the use of two categories, essentially "passing" and "failing" (1994, 54 n. 3). Similarly, Broad points out that an attempt at a contextually-sensitive scoring system which awarded 0, 3 or 6 credits to student essays failed in practice because it was not sufficiently flexible to address the "'economic realities'" of the situation; the original six-point holistic scoring guide was reinstated after one year (1994, 276 n. 9).