

Validity and Reliability Issues in the Direct Assessment of Writing

Karen L. Greenberg

During the past decade, writing assessment programs have mushroomed at American colleges and universities. Faced with legislative mandates to certify and to credential students' literacy skills, college writing teachers have become more knowledgeable about writing assessment, and they are trying to make writing tests parallel more closely their writing curricula and pedagogy.

According to national surveys of post-secondary writing assessment practices, writing teachers have developed and administered holistically-scored essay tests of writing, which they prefer over all other types of writing tests (CCCC Committee on Assessment; Greenberg, Wiener, and Donovan; Lederman, Ryzewic, and Ribaudó). These surveys confirm the growing consensus within our profession that the best way to assess students' writing skills is through writing or "direct" assessment. While there is still much disagreement about what constitutes an effective writing sample test, there is agreement that multiple-choice testing--the "indirect" assessment that once dominated post secondary writing assessment--is no longer adequate for our purposes; yet direct writing assessment continues to be challenged. This essay will elaborate on and respond to some of these challenges and will speculate on future directions in writing assessment.

The Reliability of Essay Test Scores

Most essay tests of writing are evaluated by holistic scoring, a procedure based on the response of trained readers to a meaningful "whole" piece of writing. Holistic scoring involves reading a writing sample for an overall impression of the writing and assigning the sample a score point value based on a set of scoring criteria. Typically, holistic scoring systems use a scoring scale or guide, created by composition faculty, that describes papers at different levels of competence (for examples, see Cooper; Greenberg, Wiener, and Donovan; White, 1985). In order for a holistic scoring system to be of any value, it must be shown to be "reliable," that is, it should yield the same relative magnitude of scores for the same group of

writers under differing conditions. Reliability is an estimate of a test score's accuracy and consistency, an estimate of the extent to which the score measures the behavior being assessed rather than other sources of score variance.

No test score is perfectly reliable because every testing situation differs. Sources of error inherent in any measurement situation include inconsistencies in the behavior of the person being assessed (e.g., illness, lack of sleep), variability in the administration of the test (inadequate light, insufficient space) and differences in raters' scoring behaviors (leniency, harshness). This last source of error has been the focus of almost all direct writing assessment programs, probably because it is subject to the greatest control (i.e., raters can be trained to apply scale criteria more consistently). Most programs calculate only an inter-rater scoring reliability--an estimate of the extent to which readers agree on the scores assigned to essays.

The inter-rater reliability of holistic scores, and of essay tests in general, has been under attack since 1916 when the College Entrance Examination Board (the College Board) added an hour-long essay test to its Comprehensive Examination in English. The College Board--the country's largest private testing agency and creator of multiple-choice tests of writing--has published most of these attacks, and it has done so in rather acrimonious terms: "The history of direct writing assessment is bleak. As far back as 1880 it was recognized that the essay examination was beset with the curse of unreliability" (Breland et al. 2).

During the first half of this century, essay tests did indeed have relatively low inter-rater reliability correlations. As Thomas Hopkins showed in 1921, the score that a student achieved on a College Board English exam might depend more on "which year he appeared for the examination, or on which person read his paper, than it would on what he had written" (Godshalk et al. 2). Concern with reliability of essay tests increased with the College Board's introduction of essay tests of achievement in the 1940s. In 1945, three College Board researchers examined data from various College Board essay tests and wrote a report indicating that the reliability (.58-.59) was too low to meet College Board standards (Noyes, Sale, and Stalnaker). The late 1940s and early 1950 were the heyday of indirect "component" measurement. Introductory college courses overflowed with thousands of World War II veterans who did not possess the usually expected skills and knowledge, and faculty turned to multiple-choice tests to screen and diagnose students (Ohmer 19). During the 1950s, the College Board began commissioning studies to assess the "component skills of writing ability" (Godshalk et al. 2). In 1950, Paul B. Diederich examined the correlations between the grades assigned to writers by their high school English teachers and the scores that the writers achieved on

three types of tests: multiple-choice questions on the verbal sections of the Scholastic Aptitude Test, an objective editing test, and the English Composition essay test. Diederich found that the SAT verbal score was the best predictor of the teachers' grade ("The 1950 Study"). In 1954, Edith Huddleston conducted a similar analysis of the correlations between essays written by 763 high-school students and their English grades, their scores on multiple-choice tests of composition, and their teachers' judgments of their writing ability. Huddleston concluded that the multiple-choice measures were superior to essay tests because they had higher correlations with teachers' grades.

Similar conclusions were reached by Paul Diederich and his colleagues John French and Sydell Carlton in their 1961 study of inter-rater reliability. They asked fifty-three academic and nonacademic professional writers to sort into nine piles 300 essays written by college freshmen. Using factor analysis, the researchers discovered five clusters of readers who were judging the essays on five basic characteristics: ideas, reasoning, form, flavor, and mechanics. They also found that the readers who were most influenced by one of the five characteristics also favored two or three of the other characteristics; the average inter-correlation among the five factors was .31. The researchers concluded that this low correlation was unacceptable.

However, many teachers believed that a test of writing ability should require students to write, and they found support for their belief in a comprehensive review of research in written composition done in 1963 by Richard Braddock, Richard Lloyd-Jones, and Lowell Schoer. After detailing the shortcomings in multiple-choice testing, Braddock, Lloyd-Jones, and Schoer noted flaws in several of the College Board studies described above. They pointed out that Diederich never gave his readers any standards or criteria for judging the essays, so he should not have been surprised that the readers did not often agree with each other and he should not have used this disagreement to condemn holistic scoring procedures (which usually do use scoring guides) (43). They also noted that none of Huddleston's measures included items on logic, detail, focus or clarity and that her twenty-minute essay topics did not give students an opportunity to analyze and formulate their ideas (42). In addition, Braddock, Lloyd-Jones, and Schoer commented that defenders of multiple-choice tests of writing "seem to overlook or regard as suspicious the high reliabilities that they obtained in some of their studies of essay testing" (41). They concluded with a question that still needs answering today:

In how many schools has objective testing been a good predictor of success in composition classes

precisely because those classes have emphasized grammatical and mechanical matters with little or no emphasis on central idea, analysis, organization, and content? (43)

At the same time that Braddock, Lloyd-Jones, and Schoer were writing their criticisms of multiple-choice tests, the College Board was commissioning the most comprehensive study of essay tests to date, a study that was to influence the way in which writing was assessed in America for the next two decades. Fred Godshalk, Frances Swineford, and William Coffman examined a sample of 646 high-school students, each of whom wrote five free-writing samples (two 40-minute essays and three 20-minute paragraphs) that were holistically scored by five readers. The researchers summed the scores on each of these writing samples and examined their correlations with the students' scores on multiple-choice tests and on editing tests. The results led the researchers to two major conclusions: (1) the scores on these students' multiple-choice usage and sentence construction tests correlated highly with their writing sample scores and (2) the scores on the 20-minute writing samples were as reliable as the scores on the 40-minute samples (Godshalk et al. 40-41).

These conclusions came to define writing assessment, as noted in a recent College Board publication: "This study [by Godshalk et al.] was considered for some time the quintessential study of writing assessment. The findings of this study led to the use of multiple-choice usage and sentence correction items as the primary testing devices in the College Board's English Composition Achievement Test and in other composition tests" (Breland et al. 3). Almost twenty-five years passed before the College Board commissioned another comprehensive study of direct and indirect writing assessment. During those years, multiple-choice tests continued to dominate writing assessment, but essay tests of writing began to gain popularity. In 1978, Rexford Brown, the coordinator of the first National Assessment of Educational Progress, pointed out the weaknesses in current methods of assessing writing ability. Commenting on the growing dissatisfaction of writing teachers with multiple-choice tests, he noted that their high reliability is often illusory:

Of course these tests correlate with writing ability and predict academic success; but the number of cars or television sets or bathrooms in one's family also correlate with his writing ability, and parental education is one of the best predictors there is. All existing objective tests of writing are very similar to

I.Q. tests; even the best of them can only test reading, proofreading, editing, logic, and guessing skills. They cannot distinguish between proofreading errors and process errors, reading problems and scribal stutter, failure to consider audience or lack of interest in materials manufactured by someone else.(3)

Brown also noted the invidious influences of multiple-choice tests of writing: they require a passive mental, state whereas writing samples require active mental processes, they focus on and give undue importance to the less significant components of writing (usage, spelling, and punctuation); and they are often culturally and linguistically biased (4). This last problem was documented by researchers at the California State University System, who found that multiple-choice tests of English usage produced a far more negative picture of the writing abilities of minority students than did the university's essay test (White, *Teaching and Assessing*).

By 1980, universities and colleges across the country were developing or refining their own holistically-scored essay tests of writing for varied purposes, including determining placement, diagnosing strengths and weaknesses, and certifying proficiency (CCCC Committee on Assessment; Greenberg, Weiner, and Donovan). The College Board was quick to respond to the decline in the use of multiple-choice writing assessment. In 1984, the Board commissioned Hunter Breland, Roberta Camp, Robert Jones, Margaret Morris, and Donald Rock to replicate the "Godshalk" study and to examine the reliabilities of essay and of multiple-choice measures of writing skills.

For this study, 267 students from six colleges and universities each wrote six essays on two topics in three modes (narration-description, exposition, and persuasion), and each essay was holistically scored by three readers and analyzed for errors in grammar, usage, sentence structure, and mechanics. In addition, the researchers examined students' scores on six multiple-choice tests, their course grades, and their teachers' ratings of their writing skills. Their search indicated that the essay scores had lower inter-rater reliability correlations than did the scores on the multiple-choice tests. The researchers concluded that "these problems in reliability [of essay tests] can be alleviated only through multiple essays or through the combination of essay and non-essay measures" (Breland et al. 57).

Once again the College Board was asserting that multiple-choice tests were better than essay tests because they had higher inter-rater reliability correlations. But nowhere in this College Board challenge to the reliability of essay testing--or in any of the earlier studies--was there any discussion of whether users of essay tests should strive for "perfect" reliability.

Obviously, multiple-choice scores will always be more consistent than scores on essay tests simply because machines can score a multiple-choice answer more accurately than humans can score a piece of writing. However, a "correct/incorrect" score on a multiple-choice test can never reflect the subjective, social process of writing evaluation as it genuinely occurs in the academy and in the "real world." Indeed, as Ed White recently pointed out, there can never be one "true" score for a piece of writing:

But when we evaluate student writing (not to speak of writing programs or high schools), we sometimes find differences of opinion that cannot be resolved and where the concept of the true score makes no sense. . . . Some disagreements (within limits) should not be called error, since, as with the arts, we do not really have a true score, even in theory. Yet if we imagine that we are seeking to approximate a true score, we exaggerate the negative effects of disagreement and distort the meaning of the scores we do achieve. ("Language and Reality" 192)

In other words, the differences in readers' judgments of a writing sample are often simply that--deliberate differences, not random or non random errors. Forty years ago, Stephen Wiseman, the British critic of indirect assessment, asserted this point: "It is arguable that, provided markers are experienced teachers, lack of high inter-correlation is desirable, since it points to adversity of viewpoint in the judgment of complex material . . . and the total mark gives a truer 'all-around' picture" ("Marking" 206).

Reliability is a continuum. In their selected summaries of research, Braddock, Lloyd-Jones, and Schoer cited essay test inter-rater scorer reliabilities ranging from .87 to .96 (41-42). Over the past decade, as faculty experience with scoring essay tests has grown, many colleges and universities have been able to achieve respectable inter-rater reliabilities on the scoring of their writing samples. For example, at The City University of New York (which requires a university-wide single-sample writing test that is scored holistically by two readers), the annual audits of approximately 2,000 essays per year have revealed inter-rater correlations ranging from .75 to .88 over the past seven years (Ryzewic 25). Similarly, single-sample, double-reader reliabilities on the Freshman English Equivalency Examination at the California State University and Colleges have ranged from .68 to .84 over a six year period (White, "Comparison and Contrast" 291). Administrators of writing programs need not resort to multiple-

choice testing in order to feel that students' test scores are reliable. Currently, many writing programs have essay test writing tasks and scoring criteria that eliminate sources of random error and enable readers to score reliably. Indeed, the Educational Testing Service has produced a scoring manual that enables teachers all over the country to improve the reliabilities of their essay test readings (ETS Quality Assurance Free-Response Testing Team). However, as administrators and teachers focus on improving the reliabilities of their essay tests, they must simultaneously examine their tests' validities. Stephen Wiseman's 1956 comment about validity and reliability still holds true today: "There seems to be no doubt that, over the past two or three decades, educational psychologists have slowly but steadily inflated the importance of reliability, perhaps at the expense of validity" ("Symposium" 178).

The Validity of Essay Test Scores

Validity is a controversial subject in writing assessment and in all behavioral research. Determining any test's validity involves finding evidence to establish the extent to which performance on the test corresponds to the actual behavior or knowledge that the test user wants to measure. Objections to the validity of multiple-choice writing assessment have a long history. For decades, many English teachers have claimed that multiple-choice tests of writing oversimplify and trivialize writing as the mere ability to memorize rules of grammar, spelling, and punctuation and foster instruction in memorizing discrete bits of language (Witte et al.).

The three sources of evidence for any test's validity include its content, its relationship to the underlying "construct," and its ability to predict scores on related "criterion" measures (American Psychological Association 1-4). "Content-related validity" depends on the extent to which a test reflects a specific domain of content (in this case, the content of the writing courses to which the test is attached). Many researchers have noted that there is little evidence for the content validity of multiple-choice tests of writing because their "content" is English grammar and usage; none of them samples what most teachers consider the important content of writing courses--the processes of composing, revising, and editing ideas (Brossell; Bridgeman and Carlson; Brown; CCCC Committee on Assessment; Cooper; Faigley et al.; Gere; Greenberg, Wiener, and Donovan; Lloyd-Jones; Lucas; Nystrand; Odell).

Most of the research on the validity of writing tests has focused on "criterion-related validity"--the extent to which scores on essay tests and on multiple-choice tests correlate with other measures that purport to assess

writing ability; and almost every study of criterion validity done in the past four decades has focused on the correlation between test scores and course grades. While course grade is a reasonable criterion variable for tests that are designed for admissions purposes (like the tests of the College Board and the Educational Testing Service), it is not an appropriate criterion for placement, competency, and proficiency testing. A student's grade in an English or a composition course results from many variables besides writing ability (including student motivation, attendance, and diligence).

Further, a serious problem with focusing exclusively on criterion validity is that one can easily lose sight of the skills or abilities being taught and assessed. As Rex Brown pointed out, there is a high correlation between the number of bathrooms in one's home and the grade he or she is assigned in a composition or English class, but that does not mean we should consider "quantity of bathrooms" a valid measure of writing ability.

More important than criterion validity is the evidence for a measure's "construct-related validity." In essence, construct validity is the degree to which a test score measures the psychological or cognitive construct that the test is intended to measure. To determine the construct validity of a writing test, one attempts to identify the factors underlying people's performance on the test. This is a hypothesis-testing activity, rooted in theories about the ways in which writing ability manifests itself. Because of the inadequacy of our profession's current heterogeneous theories of the nature of writing ability, however, the construct validity of most writing tests is very questionable. Only a handful of test developers have tried to analyze the domain of abilities, skills, understandings, and awareness that comprise the construct of writing ability (see Bridgeman and Carlson; Faigley et al.; Gorman, Purves, and Degenhart; and Witte for some excellent attempts at defining this construct.) Unfortunately, even the best analyses result in taxonomies of the domain, and, as Arthur Applebee has noted, "There is no widely accepted taxonomy of types of writing, and certainly none that holds up to empirical examination of the kinds of tasks on which students can be expected to perform similarly well (or poorly)"(7).

To repeat, the evidence for the construct validity of a test of writing grows out of its conceptual framework. The strategy for obtaining this evidence is to build a theoretical model of the writing process and then to examine the dimensions of the process that the test taps. This is the method that College Board researchers used in a recent attempt to provide evidence for construct validity of multiple-choice tests of writing (Breland et al.). Using factor analytic techniques, College Board researchers tested a series of hypothesized models of writing ability: a single-factor model ("general writing ability"); a three-factor model ("narrative, expository, and persuasive writing abilities"); and a hierarchical model ("general writing ability"

and "narrative, expository, and persuasive writing abilities in response to two different topics in each mode")(Breland et al. 45). They found that the model that predicted the data best was the hierarchical one, and they concluded that "while there is one dominant writing ability factor that explains about 78 percent of the common variance, there are definable subfactors based on writing topics and, to a lesser extent, mode of expression" (45).

This finding calls into question the practice of assessing writing ability with a single writing sample and also echoes recent evidence that writing ability is not a single construct but rather is a composite of several situation-specific constructs (Applebee, Langer, and Mullis; Bridgeman and Carlson; Faigley et al.; Gorman et al.; Lucas; Purves et al.). If writing ability does vary across discourse modes, the implication is that it should be assessed by two or more tasks (including tasks that require writing to construct or to communicate one's knowledge, writing to convince readers to feel or to do something, writing to entertain, and so forth). The College Board researchers, however, did not reach this conclusion, possibly because a multi-sample, multi-domain essay test is expensive, and, as they pointed out, the increase in validity might not be "cost-effective":

Although small increments in validity are possible with additional readings [of essays], the marginal cost of these additional readings is very high. Consequently, one way to reduce the cost of essay assessments is to combine them with a multiple-choice assessment. (Breland et al. 59)

However, while multiple-choice tests cost less to score than do essay tests, they are more costly to develop. Because of recent "truth-in-testing" laws, many institutions have to buy multiple forms and new revisions of their multiple-choice tests to keep them secure. Thus, in the long run, it is often less expensive to develop essay tests.

Essay tests of writing have one major advantage over multiple-choice tests: Faculty who share ideas and work together to develop an essay test often shape an exam that is grounded in their theories, curricula, and classroom practices. This has been true for teachers who have developed "portfolio" writing assessments, tests that sample several discourse domains and that provide opportunities for students to revise their writing (Anson; Belanoff and Dixon; Belanoff and Elbow; Camp, 1985; Elbow and Belanoff; Faigley et al; Lucas).

Portfolio evaluation is probably the most valid means of assessing writing available to us today because it enables teachers to assess compos-

ing and revising across a wide range of communicative contexts and tasks (Camp, "The Writing Folder"). Moreover, this process sends the message that the construct of "writing" means developing and revising extended pieces of discourse, not filling in blanks in multiple-choice exercises or on computer screens. It communicates to everyone involved--students, teachers, parents, and legislators--our profession's beliefs about the nature of writing and about how writing is taught and learned.

Further, multi-sample writing tests enable teachers to evaluate parts of the writing processes that so many of us emphasize in our curricula. As Leo Ruth and Sandy Murphy have pointed out, it is possible to design large-scale writing tests that preserve more of the steps of "real" writing than normally occur in most testing situations (113-114). For example, in England's version of a national assessment of writing, teachers administering the test introduce the writing task and discuss it before students begin writing. In addition, three samples of writing, each involving different types of tasks, are collected (113). In Ontario, Canada, students have two days to take their writing assessment. On the first day, students generate and record ideas; and on the second day, they write and revise their essays (114).

Multi-sample portfolio tests seem more relevant to our theories about the construct of writing and to our classroom practices than do other writing assessment measures. However, questions about the scoring of these tests remain unresolved. Is holistic scoring the most appropriate method of rating writing samples? Are holistic ratings based on the consistent application of mutually agreed-upon substantive criteria for "good writing"? Researchers have not addressed these questions extensively, and the results of existing research are contradictory. Several studies indicated that holistic scores correlate strongly with "superficial" aspects of writing, such as handwriting, spelling, word choice, and errors (Greenberg, 1981; McColly; Nold and Freedman; Neilson and Piche). However, the essays being scored in these studies were quite brief (written in time periods ranging from twenty to sixty minutes), and one can argue that readers were predisposed to attend to these salient but superficial errors. To my knowledge, no one has yet analyzed the holistic scores of raters who evaluate tests that require writers to do extensive revision or to write multi-sample portfolios.

The question of substantive criteria for "good writing" relates directly to the issue of construct validity. Lacking a theoretical model of effective writing ability, most test developers fall back on descriptions of text characteristics for inclusion as criteria in holistic scoring guides. An examination of these guides reveals that many of them are quite similar, indicating some professional consensus about the characteristics of effective

writing across a wide variety of tasks and testing purposes. Nevertheless, the skills described in the criteria on current holistic scoring guides do not provide an adequate definition of "good writing" or of the many factors that contribute to effective writing in different contexts. Criteria on holistic scoring guides cannot accommodate many of the cognitive and social determinants of effective writing, not the least of which include the writer's intentions and the situational expectations of the writer and of potential readers. Nor can holistic criteria assess a writer's composing processes or the ways in which these processes--and products--vary in relation to the writer's purpose, audience, role, topic, and context. But should they?

Evidence for validity depends on the purpose for which the assessment instrument is used. According to the surveys cited earlier, most post-secondary writing assessment in America is done for entry-level placement purposes (CCCC Committee on Assessment; Greenberg, Wiener, and Donovan; Lederman, Ryzewic, and Ribaud). The second most frequently cited purpose is to certify that students have mastered the competencies that they practiced in a specific writing course. These surveys revealed that the majority of the respondents whose schools used holistically-scored essay tests for these purposes were "very satisfied" with the results of the tests. These findings indicate that many users of holistically-scored essay tests believe that the tests can provide adequate and appropriate information about students' abilities to function successfully in different types of writing courses. In addition, holistically-scored essay tests may provide a more accurate assessment of the writing skills of minority students than multiple-choice testing can provide. The California State University research (mentioned earlier) indicated that many more Black, Mexican-American, and Asian-American students than white students received the lowest possible score on a multiple-choice test of usage but earned a middle or high score from trained essay readers for their writing ability (White, *Teaching and Assessing*). This disparity casts further doubt on the validity of multiple-choice testing as an indicator of writing ability.

Research is needed to determine whether holistically-scored portfolio tests can provide diagnostic information and can function as fully contextualized assessments of students' writing competence or improvement. As Brian Huot pointed out in his recent review of research on holistic scoring, "We need to question and explore the particular problems associated with the specific uses of holistic scoring" (208). What we have now, holistically-scored essay tests, serve our limited purposes very well. What we still need--a multi-draft instrument that adequately represents writing in different discourse domains for different purposes and for different discourse communities--is an inchoate vision that many of us share.

The Implications of These Challenges for Our Profession

Writing teachers are asked to do more assessing than are any other humanities colleagues, yet many of us are particularly subject to insecurity about our ability to understand and manipulate data. It is no coincidence that most of the research on writing and writing assessment that followed the 1966 publication of the College Board's "Godshalk" study borrowed the quantitative empiricism of research in the physical sciences. This reverence for objective data diminished in the early 1980s, partially in response to the publication of Janet Emig's contextualized research and her scathing indictments of empirical and experimental research. One of the recent outgrowths of the trend toward contextualized research on writing was a consensus about the need for naturalistic, context-rich, qualitative models of evaluating students' writing. Current portfolio measures come closest to capturing these models.

Those of us who are committed to the direct assessment of writing understand that we do not have to model our programs on multiple-choice assessments, that there is no need to create the "perfect" essay test. Readers will always differ in their judgments of the quality of a piece of writing; there is no one "right" or "true" judgment of a person's writing ability. If we accept that writing is a multidimensional, situational construct that fluctuates across a wide variety of contexts, then we must also respect the complexity of teaching and testing it.

Works Cited

- American Psychological Association. *Joint Technical Standards for Educational and Psychological Testing*, 4th Draft. Washington, D.C.: American Psychological Association, 1984.
- Anson, Chris. "Portfolio Assessment Across the Curriculum." *Notes from the National Testing Network in Writing* 8 (1988): 6-7.
- Applebee, Arthur. "Musings." *Research in the Teaching of English* 22 (1988): 6-8.
- Applebee, Arthur, Judith Langer, and Ina Mullis. *The Writing Report Card: Writing Achievement in American Schools*. Princeton, NJ: Educational Testing Service, 1987.
- Belanoff, Patricia, and Marcia Dixon. *Portfolio Grading: Process and Product*.

Portsmouth, NH: Heinemann, 1991.

- Belanoff, Patricia, and Peter Elbow. "Using Portfolios to Increase Collaboration and Community in a Writing Program." *WPA: Writing Program Administration* 9.3 (1986): 27-40.
- Braddock, Richard, Richard Lloyd-Jones, and Lowell Schoer. *Research in Written Composition*. Champaign, IL: National Council of Teachers of English, 1963.
- Breland, Hunter M., et al. *Assessing Writing Skill*. New York: College Entrance Examination Board, 1987.
- Bridgeman, Brent, and Sybil Carlson. *Survey of Academic Writing Tasks Required of Graduate and Undergraduate Foreign Students*. Princeton, NJ: Educational Testing Service, 1983.
- Brossell, Gordon. "Current Research and Unanswered Questions in Writing Assessment." *Writing Assessment: Issues and Strategies*. Eds. Karen Greenberg, Harvey Wiener, and Richard Donovan. New York: Longman, 1986. 168-182.
- Brown, Rexford. "What We Know Now and How We Could Know More About Writing Ability in America." *Journal of Basic Writing* 4 (1978): 1-6.
- Camp, Roberta. "Thinking Together About Portfolios." *The Quarterly of the National Writing Project* 12 (1990): 8-14.
- Camp, Roberta. "The Writing Folder in Post-Secondary Assessment." *Directions and Misdirections in English Evaluation*. Ed. Peter J.A. Evans. Ottawa, Canada: Canadian Council of Teachers of English, 1985. 91-99.
- CCCC Committee on Assessment. *Post-secondary Writing Assessment: An Update on Practices and Procedures*. (Spring 1988). Report to the Executive Committee of the Conference on College Composition and Communication.
- Cooper, Charles. "Holistic Evaluation of Writing." *Evaluating Writing: Describing, Measuring, Judging*. Eds. Charles Cooper and Lee Odell. Urbana, IL: National Council of Teachers of English, 1977. 3-32.
- Diederich, Paul B. "The 1950 College Board English Validity Study." *Research Bulletin RB 50-58*. Princeton, NJ: Educational Testing Service, 1950.
- . *Measuring Growth in English*. Urbana, IL: National Council of

- Teachers of English, 1974.
- Diederich, Paul B., John W. French, and Sydell T. Carlton. "Factors in Judgments of Writing Ability." *Research Bulletin RB 61-15*. Princeton, NJ: Educational Testing Service, 1961.
- Elbow, Peter, and Pat Belanoff. "Portfolios as Substitutes for Proficiency Examinations." *College Composition and Communication* 37 (1987): 336-339.
- Emig, Janet. *The Composing Processes of Twelfth Graders*. Urbana, IL: National Council of Teachers of English, 1971.
- ETS Quality Assurance Free-Response Testing Team. *Guidelines for Developing and Scoring Free-Response Testing*. Princeton, NJ: Educational Testing Service, 1987.
- Faigley, Lester et al. *Assessing Writers' Knowledge and Processes of Composing*. Norwood, NJ: Ablex, 1985.
- Gere, Ann. "Written Composition: Toward a Theory of Evaluation." *College English* 42 (1980): 44-58.
- Godshalk, Fred et al. *The Measurement of Writing Ability*. New York: College Entrance Examination Board, 1966.
- Gorman, Tom, Alan Purves, and Elaine Degenhart, Eds. *The IEA Study of Written Composition*. Vol. 1. New York: Pergamon Press, 1988.
- Greenberg, Karen. "Competency Testing: What Role Should Teachers of Composition Play?" *College Composition and Communication* 33 (1982): 366-378.
- . *The Effects of Variations in Essay Questions on the Writing Performance of CUNY Freshmen*. New York: CUNY Instructional Resource Center, 1981.
- Greenberg, Karen, Harvey Wiener, and Richard Donovan. "Preface." *Writing Assessment: Issues and Strategies*. Eds. Karen Greenberg, Harvey Wiener, and Richard Donovan. New York: Longman, 1986. xi-xvii.
- Huddleston, Edith M. "Measurement of Writing Ability at the College Level: Objective vs. Subjective Techniques." *Journal of Experimental Education* 22 (1954): 165-213.
- Huot, Brian. "Reliability, Validity, and Holistic Scoring: What We Know and What We Need to Know." *College Composition and Communication* 41 (1990): 201-213.
- Lederman, Marie Jean, Susan Ryzewic, and Michael Ribaud. *Assessment and Improvement of the Academic Skills of Entering Freshmen: A National Survey*. New York: CUNY Instructional Resource Center, 1983.
- Lloyd-Jones, Richard. "Skepticism about Test Scores." *Notes from the National Testing Network in Writing* 3 (1983): 9.
- Lucas, Catharine Keech. "Recontextualizing Literacy Assessment." *The Quarterly* 10 (1988): 4-10.
- McColly, William. "What Does Educational Research Say about the Judging of Writing Ability?" *Journal of Educational Research* 64 (1970): 147-156.
- Neilson, Lorraine, and Gene Piche. "The Influence of Headed Nominal Complexity and Lexical choice on Teachers' evaluation of Writing." *Research in the Teaching of English* 15 (1981): 65-74.
- Nold, Ellen, and Sarah Freedman. "An Analysis of Readers' Responses to Essays." *Research in the Teaching of English* 11 (1977): 164-173.
- Noyes, Edward S., William M. Sale, and John M. Stalnaker. *Report on the First Six Tests in English Composition*. New York: College Entrance Examination Board, 1945.
- Nystrand, Martin. "An Analysis of Errors in Written Communication." *What Writers Know*. Ed. Martin Nystrand. New York: Academic Press, 1982. 57-73.
- Odell, Lee. "Defining and Assessing Competence in Writing." *The Nature and Measurement of Competency in English*. Ed. Charles Cooper. Urbana, IL: National Council of Teachers of English, 1981. 95-138.
- Purves, Alan et al. "Towards a Domain-Referenced System for Classifying Composition Assignments." *Research in the Teaching of English* 18 (1984): 417-438.
- Ruth, Leo, and Sandra Murphy. *Designing Tasks for the Assessment of Writing*. Norwood, NJ: Ablex, 1988.
- Ryzewic, Susan. *The CUNY Writing Assessment Test: A Three-Year Audit*. New York: CUNY Instructional Resource Center, 1982.

White, Edward, ed. *Comparison and Contrast: The California State University Freshman English Equivalency Examination*. Vol. 8. Long Beach: California State University, 1981.

_____. "Language and Reality in Writing Assessment." *College Composition and Communication* 41 (1990): 187-200.

_____. *Teaching and Assessing Writing*. San Francisco: Jossey-Bass, 1985.

Wiseman, Stephen. "The Marking of English Composition in Grammar School Selection." *British Journal of Educational Psychology* 19 (1949): 200-209.

_____. "Symposium: The Use of Essays on Selection of 11+." *British Journal of Educational Psychology* 26 (March 1956): 172-179.

Witte, Stephen. "Direct Assessment and the Writing Ability Construct." *Notes from the National Testing Network in Writing* 8 (1988): 13-14.

Witte, Stephen, et al. "Literacy and the Direct Assessment of Writing: A Diachronic Perspective." *Writing Assessment: Issues and Strategies*. Eds. Karen Greenberg, Harvey Wiener, and Richard Donovan. New York: Longman, 1986. 13-34.



MAYFIELD PUBLISHING COMPANY

For your second-semester courses—

RESPONDING TO LITERATURE

Judith A. Stanford, Rivier College

Responding to Literature encourages the reader's personal response to a large and richly diverse selection of literature, including the essay. Three introductory chapters use student papers to illustrate ways of responding to and writing about literature.
Paperbound / 1580 pages

THE CRITICAL READER, THINKER, AND WRITER

W. Ross Winterowd and Geoffrey R. Winterowd, University of Southern California

This text-reader makes accessible to students a variety of powerful approaches to critical reading, thinking, and writing through 51 readings diverse as to discipline, cultural background, and genre.
Paperbound / 608 pages

MOTIVES FOR WRITING

Robert Keith Miller, University of St. Thomas

Suzanne S. Webb, Texas Woman's University

Following a lucid introduction to the rhetorical situation and the writing process, 74 readings show how writers from diverse backgrounds have realized the most common motives for writing.
Paperbound / 608 pages

New, for advanced composition—

DEVELOPING A WRITTEN VOICE

Dona Hickey, University of Richmond

This practical text on voice and style firmly roots students in traditional rhetoric while inductively encouraging them to explore important stylistic issues within a collaborative context.
Paperbound / 228 pages

For creative writing—

WORKING WORDS: THE PROCESS OF CREATIVE WRITING

Wendy Bishop, Florida State University

This multigenre introduction to creative writing focuses on the process of creative writing before turning to genre distinctions and end products.
Paperbound / 336 pages

MAYFIELD PUBLISHING COMPANY

1240 VILLA STREET • MOUNTAIN VIEW, CA 94041

For orders or inquiries, please telephone (800)433-1279 or FAX (415)960-0328.