ED 335 681 CS 212 958

AUTHOR	Freedman, Sarah Warshauer
TITLE	Evaluating Writing: Linking Large-Scale Testing and Classroom Assessment. Occasional Paper No. 27.
INSTITUTION	Center for the Study of Writing, Berkeley, CA.; Center for the Study of Writing, Pittsburgh, PA.
SPONS AGENCY	Office of Educational Research and Improvement (ED), Washington, DC.
PUB DATE	May 91
NOTE	25p.
PUB TYPE	Reports - Descriptive (141)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Elementary Secondary Education; Foreign Countries; Informal Assessment; *Portfolios (Background Materials); Program Descriptions; Test Format; *Testing Problems; Test Length; *Test Reliability; Test Validity; *Writing Evaluation; *Writing Tests
IDENTIFIERS	College Entrance Examination Board; Educational Testing Service; National Assessment of Educational Progress; United Kingdom; Vermont

ABSTRACT

Writing teachers and educators can add to information from large-scale testing and teachers can strengthen classroom assessment by creating a tight fit between large-scale testing and classroom assessment. Across the years, large-scale testing programs have struggled with a difficult problem: how to evaluate student writing reliably and cost-effectively. Indirect measures, direct assessments, "holistic" scoring, and primary trait scoring (as used by the College Entrance Examination Board, the Educational Testing Service, and the National Assessment of Educational Progress) all have serious limitations. Even though not well defined, the portfolio movement provides a potential link between large-scale testing and classroom assessment and teaching. Several large-scale portfolio assessment programs are currently in place: (1) the Arts PROPEL program, a Pittsburgh school-district portfolio project in art, music, and imaginative writing; (2) the "Primary Language Record," a kind of portfolio introducing systematic record-keeping about language growth into all elementary classrooms in the United Kingdom; (3) a draft, state-wide plan for portfolio assessment in Vermont; and (4) the General Certificate of Secondary Education (GSCE) in language and literature, in which British students choose either a timed writing test plus a portfolio of coursework or simply a folder of coursework. However, just collecting and evaluating portfolios will solve neither the assessment problems nor the need to create a professional climate in schools. By coupling assessment and instruction in increasingly sophisticated ways, educators and teachers may be able to make a real difference in education. (Seventy-three references are attached.) (RS)

Center for the Study of Writing

Occasional Paper No. 27

EVALUATING WRITING: LINKING LARGE-SCALE TESTING AND CLASSROOM ASSESSMENT

Sarah Warshauer Freedman

May, 1991

University of California, Berkeley
Carnegie Mellon University

CENTER FOR THE STUDY OF WRITING

Occasional Paper No. 27

EVALUATING WRITING: LINKING LARGE-SCALE TESTING AND CLASSROOM ASSESSMENT

Sarah Warshauer Freedman

May, 1991

To appear in *Restructuring Learning: Analysis and Recommendations of the Council of Chief State School Officers.* Washington, DC: Council of Chief State School Officers.

University of California
Berkeley, CA 94720

Carnegie Mellon University
Pittsburgh, PA 15213

The publication of this report was funded by grant number R117G00036 from the Office of Educational Research and Improvement, U.S. Department of Education (OERI/ED) for the National Center for the Study of Writing and Literacy. However, the opinicns expressed herein do not necessarily reflect the position or policy of the OERI/ED and no official endorsement by the OERI/ED should be inferred.

This publication was produced on an Apple Macintosh IIsi computer with portrait display monitor and an Apple LaserWriter IIntx printer donated to the National Center for the Study of Writing and Literacy by Apple Computer, Inc.

CENTER FOR THE STUDY OF WRITING

EVALUATING WRITING: LINKING LARGE-SCALE TESTING AND CLASSROOM ASSESSMENT

Sarah Warshauer Freedman
University of California at Berkeley

Robert Hogan, then executive director of the National Council of Teachers of English, opens his preface to Paul Diederich's 1974 book *Measuring Growth in English* with the following words:

> Somehow the teaching of English has been wrenched out of the Age of Acquarius and thrust into the Age of Accountability. Many of us view educational accountants in much the same spirit as we view the agent of the Internal Revenue Service coming to audit our returns. Theoretically, it is possible the agent will turn out to be a pleasant person, gregarious and affable, who writes poetry in his free time and who will help us by showing how we failed to claim all our allowable deductions, so that the result of the audit is the discovery of a new friend and a substantial refund. But somehow we doubt that possibility.
>
> For the specialist in measurement and testing we have our image, too. In his graduate work, one of the foreign languages he studied was statistics. And he passed it. The other one was that amazing and arcane language the testing specialists use when they talk to one another. He passed it, too, and is fluent in it. He doesn't think of children except as they distribute themselves across deciles. He attempts with his chi-scuares to measure what we've done without ever understanding what we were trying to do. (p. iii)

Most English teachers, I suspect, would still agree with Hogan's remarks. I will focus in this paper on bridging this rather wide gap between teachers of writing and the testing and measurement community. I will focus on two currently distinct kinds of writing evaluation—*large-scale testing* at the national, state, district, and sometimes school levels, the natural domain of the educational accountants, and *classroom assessment* by teachers looking at their own students inside their own classrooms, teachers who see kids and not distributions of deciles but whose judgments, according to measurement specialists, may be unreliable and biased.[1] In writing, as in most areas of the curriculum, large-scale testing and classroom assessment normally serve different purposes and quite appropriately assume different forms. However, if we could create a tight fit between large-scale testing and classroom assessment, we could potentially add to the kinds of information we now get from large-scale testing programs, and we could help teachers strengthen their classroom assessments and thereby their teaching and their students' learning.

[1] In this paper the term testing will refer to large-scale standardized evaluation and assessment will refer to the evaluative judgments of the classroom teacher. Calfee (1987) describes testing activities as usually "group administered, multiple choice, mandated by external authorities, used by the public and policy makers to decide 'how the schools are doing'" while assessment activities include "evaluation of individual student performance, based on the teacher's decisions about curriculum and instruction at the classroom level, aimed toward the student's grasp of concepts and mastery of transferrable skills (Calfee and Drum, 1979)" (p. 738).

Before presenting some ideas for linking large-scale testing and classroom assessment, I will provide background about the form of most large-scale writing tests and will discuss their limitations. I will then describe portfolio assessment, an important innovation in classroom writing evaluation that is filtering up in some cases to the state level and now even to the National Assessment of Educational Progress (NAEP). Portfolio assessment contains the foundations for potential formal links between large-scale testing and classroom assessment levels. Finally, I will give several examples of portfolio programs at work, examples that I find helpful as I think about possible future directions for writing assessment and instruction in this country: a large-scale, classroom-centered portfolio effort for elementary students in England, *The Primary Language Record*; a state-level portfolio assessment from Vermont for grades 4 and 11; and a large-scale national examination at the secondary level in Great Britain, the General Certificate of Secondary Education (GCSE).

Large-scale Testing

Historically, the large-scale testing of writing has developed to fulfill a number of purposes: (a) to certify that students have mastered writing at some level (e.g., the National Assessment of Educational Progress); (b) to evaluate writing programs in the school, district, or in some cases classroom (e.g., the California Assessment Program); (c) to place students in programs or classes (e.g., many college-level placement examinations given to freshmen); (d) to decide the fate of individuals with respect to admissions, promotion, or graduation ("gatekeeping") (e.g., the SAT, high school graduation tests, writing samples gathered by potential employers). Unlike classroom assessment, large-scale testing generally has not been concerned with charting the development of individual writers.

Across the years, large-scale testing programs have struggled with a difficult problem: how to evaluate student writing reliably and cost-effectively. One highly criticized but commonly used way is through *indirect* measures designed to provide proxies for writing abilities. Indirect measures are generally multiple-choice tests and typically include questions about grammar or sentence structure or scrambled paragraphs to be rearranged in a logical order. These indirect measures are in widespread use; in 1984, 19 states measured writing indirectly while only 13 had direct measures, and 18 had no measures at all (Burstein et al., 1985, in Baker, 1989). The appeal of indirect measures of writing is obvious; they're quick to administer and cheap to score. The problems are obvious too; indirect measures are poor predictors of how well the test-taker actually writes. According to Gertrude Conlan (1986), long-time specialist in writing assessment at Educational Testing Service:

> No multiple-choice question can be used to discover how well students can express their own ideas in their own words, how well they can marshal evidence to support their arguments, or how well they can adjust to the need to communicate for a particular purpose and to a particular audience. Nor can multiple-choice questions ever indicate whether what the student writes will be interesting to read. (p. 124)

And if we believe Resnick and Resnick (1990) that "[y]ou get what you assess," multiple-choice writing tests will have negative effects on instruction since teaching to the test would not include asking students to write.

From 1890 on into the 1960s the College Entrance Examination Board (CEEB) struggled to find practical ways to move away from multiple-choice, indirect measures of writing. The goal was to design *direct* assessments that would include the collection and scoring of actual samples of student writing (Diederich, French, & Carlton, 1961; Godshalk, Swineford, & Coffman, 1966; Huddleston, 1954; Meyers, McConville, & Coffman, 1966). CEEB's struggles were many. First of all, the student writing would have to be evaluated. Besides the expense of paying humans to score actual writing samples, it proved difficult to get them to agree with one another on even a single general-impression score. In 1961 Diederich, French, and Carleton at the Educational Testing Service (ETS) conducted a study in which "sixty distinguished readers in six occupational fields" read 300 papers written by college freshmen (in Diederich, 1974, p. 5). Of the 300 papers, "101 received every grade from 1 to 9" (p. 6). On as many papers as they could, the readers wrote brief comments about what they liked and disliked. These comments helped ETS researchers understand why readers disagreed.

During the 1960s ETS and the CEEB developed ways of training readers to agree independently on "holistic" or general impression scores for student writing, thus solving the reliability problems of direct assessment (Cooper, 1977; Diederich, 1974). For this scoring, readers are trained to evaluate each piece of student writing relative to the other pieces in the set, without consideration of standards external to the examination itself (Charney, 1984). Besides figuring out how to score the writing reliably, the testing agencies also figured out ways to collect writing samples in a controlled setting, on assigned topics, and under timed conditions. With the practical problems solved and routines for testing and scoring in place, the door opened to the current, widespread, large-scale, direct assessments of writing (Davis, Scriven, & Thomas, 1987; Diederich, 1974; Faigley et al., 1985; Myers, 1980; White, 1985).

When direct writing assessments were relatively novel, the profession breathed a sigh of relief that writing could be tested by having students write. Diederich's opening to his 1974 book typified the opinions of the day:

> As a test of writing ability, no test is as convincing to teachers of English, to teachers in other departments, to prospective employers, and to the public as actual samples of each student's writing, especially if the writing is done under test conditions in which one can be sure that each sample is the student's own unaided work. (p. 1)

However, Diederich's words sound dated now. With large-scale direct assessments of writing in widespread use, educators are already raising questions about their validity, just as they did and continue to do for the indirect measures provided by multiple-choice tests. Many tensions center around the nature of test-writing itself. Although controlled and written under unaided conditions, as Diederich points out, such writing has little function for students other than for them to be evaluated. Too, students must write on topics they have not selected and may not be interested in. Further, they are not given sufficient time to engage in the elaborated processes that are fundamental to how good writers write and to how writing ideally is taught (Brown, 1986; Lucas, 1988a,b; Simmons, 1990; Witte et al., in press). In short, the writing conditions are "unnatural." Finally, educators often make claims about writing in general and students' writing abilities based on one or perhaps a few kinds of writing, written in one kind of context, the testing setting.

Current debates surrounding the NAEP writing assessment provide important illustrations of the tensions surrounding most large-scale, direct writing assessments. The goal of the NAEP assessment is to provide at five-year intervals "an overall portrait of the writing achievement of American students in grades 4, 8, and 11" (1990b, p. 9) as well as to mark changing "trends in writing achievement" across the years (1986a, p. 6). The National Assessment gathers informative, persuasive, and imaginative writing samples from students at the three grade levels. For eighth- and twelfth-graders, the test "is divided into blocks of approximately 15 minutes each, and each student is administered a booklet containing three blocks as well as a six-minute block of background questions common to all students" (1986a, p. 92). During a 15-minute block, students write on either one or two topics. For fourth-graders, the blocks last only 10 minutes (National Assessment of Educational Progress, 1990a). This means that fourth-graders have had between 5 and 10 minutes to produce up to four pieces of writing during a 30-minute test; eighth- and twelfth-graders have had between 7 1/2 and 15 minutes to produce up to four pieces during a 45-minute test (National Assessment of Educational Progress, 1990a).

For good reason, writing researchers and educators have critiqued the National Assessment, arguing that it is not valid to make claims about the writing achievement of our nation's schoolchildren given the NAEP testing conditions, especially the short time students have for writing, and given the way the writing is evaluated (e.g., see Mellon, 1975; Nold, 1981; Silberman, 1989). With respect to the testing conditions, the NAEP report writers themselves caution:

> The samples of writing generated by students in the assessments represent their ability to produce first-draft writing on demand in a relatively short time under less than ideal conditions; thus, the guidelines for evaluating task accomplishment are designed to reflect these constraints and do not require a finished performance. (1990b, p. 7)

Based on NAEP writing data, how confident can we be in the following claim made in *The Writing Report Card*: "A major conclusion to draw from this assessment is that students at all grade levels are deficient in higher-order thinking skills" (1986a, p. 11)? Can students possibly reveal their higher-order thinking skills in 15 minutes when writing on an assigned topic that they have never seen?

In stark contrast to most testing conditions and consistent with our sense of how writing can be used to support the development of sophisticated higher order thinking, the pedagogical and research literature in writing from the past decade shows that higher-order thinking occurs when there is an increased focus on a writing process which includes encouraging students to take lots of time with their writing, to think deeply and write about issues in which they feel some investment, and to make use of plentiful response from both peers and teachers as they revise (Dyson & Freedman, in press; Freedman, 1987). Most tightly timed test-type writing goes against current pedagogical trends. What Mellon (1975) pointed out about the NAEP writing assessment some 15 years ago remains true today:

> One problem with the NAEP essay exercises, which is also a problem in classroom teaching, is that the assessors seem to have underestimated the arduousness of writing as an activity and consequently overestimated the level of investment that unrewarded and unmotivated students would bring to the task. After all, the students were asked to write by examiners whom they did not know. They were told that their teachers would not see their writing, that it would not influence their marks or academic futures, and presumably that they would receive no feedback at all on their efforts.
>
> Clearly this arrangement was meant to allay the students' fears, but its effect must have been to demotivate them to some degree, though how much is anyone's guess. We all know that it is difficult enough to devote a half hour's worth of interest and sustained effort to writing externally imposed topics carrying the promise of teacher approbation and academic

> marks. But to do so as a flat favor to a stranger would seem to require more generosity and dutiful compliance than many young people can summon up.
>
> . . . Answering multiple choice questions without a reward in a mathematics assessment or a science lesson may be one thing. Giving of the self what one must give to produce an effective prose discourse, especially if it is required solely for purposes of measurement and evaluation, is quite another. (p. 34)

NAEP is attempting to respond to these criticisms about the time for the testing. In 1988 NAEP gave a subsample of the students twice as much time on one informative, persuasive, and imaginative topic at each grade level (20 minutes for grade 4 and 30 minutes for grades 8 and 12) (National Assessment of Educational Progress, 1990a). The results show that with increased time all students scored significantly better on the narrative tasks and fourth- and twelfth- graders scored significantly better on the persuasive tasks; only the informative tasks showed no differences. Most disturbing, the extra time proved more helpful to White students than to Blacks or Hispanics, widening the gaps between these groups in the assessment results.

For the 1992 assessment, NAEP plans to provide more time across the board:

> As a result of both the findings from this study and the desire to be responsive to the latest developments in writing instruction and assessment, the response time will be increased for all writing tasks administered in the 1992 NAEP assessment. At grade 4, students will be given 25 minutes to perform each task, and at grades 8 and 12, students will be given either 25 or 50 minutes. These tasks will be designed to encourage students to allocate their time across various writing activities from gathering, analyzing, and organizing their thoughts to communicating them in writing. (1990a, p. 87)

Providing 25 or even 50 minutes for writing on a given topic will probably prove insufficient to quiet NAEP critics, since even that amount of time will not resolve the basic discrepancy between what we argue should be happening in classrooms and what happens in this testing setting. Furthermore, the findings about Blacks and Hispanics raises a new set of questions about equity and testing, not to mention equity in classroom opportunities to learn. Besides the double time, NAEP is also collecting portfolios of student writing produced as a natural part of writing instruction. The assessors have not yet decided how to evaluate the portfolios, but these data promise to provide important supplementary information for the Assessment. It will be important to remember that as the Assessment changes, the only way to collect data about trends across time will be to keep some parallel tasks. Thus, 15-minute samples will still be used for the trend studies and conclusions about trends will be based on these very short samples.

Another major point of tension in the National Assessment centers around the issue of scoring. In an effort to obtain more information than a single holistic score and to define clearly the features of writing being judged, in the mid 1970s NAEP developed an additional scoring system, "the Primary Trait Scoring method" (Lloyd Jones, 1977, p. 33). While the criteria for judging writing holistically emerge from the writing the students do, the goal of primary trait scoring is to set specific criteria for successful writing on a particular topic ahead of time. The primary trait is determined and defined by the test-maker who decides what will be essential to writing successfully on each topic on the test. Traits vary depending on the topics. Tensions arise because the test-makers cannot always anticipate precisely what test-takers will do to produce good writing on a particular topic, and what is primary or whether one aspect of writing should be labeled primary is a debatable point.

The dilemmas come across clearly through an analysis of Lloyd Jones's (1977) example of a primary trait scoring rubric. Lloyd Jones explains that in one NAEP prompt children were to write about the following: "Some people believe that a woman's place is in the home. Others do not. Take ONE side of this issue. Write an essay in which you state your position and defend it" (p. 60). The directions for scoring this trait show the conflicts that are likely to emerge between a primary trait and a holistic score representing the general quality of the student's writing. The writing receives a 0 score if the writer gives no response or a fragmented response; it receives a 1 score if the writer does not take a clear position, takes a position but gives no reason, restates the stem, gives and then abandons a position, presents a confused or undefined position, or gives a position without reasons; it receives a 2 if the writer takes a position and gives one unelaborated reason; it receives a 3 if the writer takes a position and gives one elaborated reason, one elaborated reason plus one unelaborated reason, or two or three unelaborated reasons; it receives a 4 if the writer takes a position and gives two or more elaborated reasons, one elaborated reason plus two or more unelaborated reasons, or four or more unelaborated reasons.

What happens to the student who does not follow directions to take "ONE" position on a woman's place but points out the complexity of the issue rather than taking a side, perhaps showing how a woman has many places, in the home and out? This student would receive a 1 score but might write a substantially better essay than a student who receives a 2, 3, or 4 score for taking a side and providing one or more reasons. In another scenario a student who gives one elaborated reason for a 3 score could write a far better essay than the student who gives four or more unelaborated reasons and receives a 4. NAEP scoring rubrics seem to have gotten less specific and therefore less controversial over the years.

Besides these issues of judging elaboration particular to this scoring rubric, the primary trait score only measures one aspect of writing. By contrast, a holistic score takes into account the whole piece—including its fluency, sentence structure, organization, coherence, mechanics, and idea development. Indeed, in a study comparing holistic and primary trait scoring, NAEP found that primary trait scoring does not correlate particularly well with holistic quality judgments; correlations ranged from .38 to .66 depending on the topic (1986a, p. 84). Freedman (1979) found that holistic scores are based primarily on how well writers develop their ideas and then organize them, but once writers do a good job at development and organization, then the rater counts syntax and mechanics.

Whereas NAEP uses a holistic score, a primary trait score, and a mechanics score for its trends reports (1986b, 1990b), NAEP uses only primary trait scoring for the reports on the status of writing for a given year (1986a, 1990a). In the latest status report, NAEP (1990a) explains, "The responses were not evaluated for fluency or for grammar, punctuation, and spelling, but information on these aspects of writing performance is contained in the writing trend report" (p. 60).

At the state level the issues in large-scale, direct, writing assessment are similar to those illustrated by the debates surrounding NAEP. States with direct writing assessments are facing the same challenges as NAEP, and several states are meeting the challenges in interesting ways. For example, let's look at the case of Alaska (Calkins, personal correspondence). Two years ago in an effort to increase accountability the Alaska state school board mandated the Iowa Test of Basic Skills for grades four, six, and eight. The Iowa test, developed in 1929, contains multiple-choice items in grammar and sentence structure, but the introduction to the test explicitly says that it is not designed to test writing skills. Alaska teachers of writing are well organized through the Alaska Writing Consortium, an affiliate of the National Writing Project, and with strong leadership in the State Department of Education. Open to the accountability concerns of the State Board and anxious to learn about the fruits of their classroom efforts, Consortium members proposed a direct writing assessment that would yield information about students' writing achievement beyond whatever other information the Iowa test might provide. The state funded an experiment at the tenth-grade level, and in 1989-1990 twelve districts participated voluntarily. The writing was scored with an analytic scale, the third method besides primary trait, and holistic scoring that is commonly used in large-scale, direct writing assessments. The analytic scale offers more information than a single holistic score but avoids some of the problems associated with primary trait scoring.[2] The analytic scale differs from primary trait because the categories are generic to good writing and are thus independent of a given topic. On this scale raters give separate scores on ideas, organization, wording, flavor, usage and sentence structure, punctuation and other mechanics, spelling, and handwriting (Diederich, 1974). An analytic scale is used by the International Association for the Evaluation of Educational Achievement (IEA) studies of written language (Gorman et al., 1988; Gubb et al., 1987).

For the Alaska test, teachers also wanted to maintain some control over the testing conditions while allowing students more natural and comfortable writing conditions than is usual for large-scale, formal assessments. Thus, students were given a common prompt but were allowed twc 50-minute time blocks on separate days to complete the writing. For the Alaska experiment 60 papers from each of the districts were scored, enough writing to provide a substantial amount of information about student writing beyond what the state board could get from the Iowa test that they were using. In particular the direct testing showed that knowledge of sentence structure does not guarantee good ideas. The board also learned that direct assessments were easy to administer and cost-effective. This past year 22 districts out of Alaska's 54 districts volunteered to participate, and Alaska teachers are experimenting with other assessment alternatives as well. To these alternatives, emerging mostly from the classroom up, I will now turn.

New Directions: Writing Portfolios

The portfolio movement provides a potential link between large-scale testing and classroom assessment and teaching, and could serve as an impetus for important reforms on all fronts, bringing together Hogan's accountants or IRS agents and the teachers whom they audit. Mostly classroom-based and designed to provide information about student growth, portfolios really are not much more than collections of student writing. They have long been a staple of many informal classroom assessments marked by careful teacher observation and careful record keeping (e.g., anecdotal records, folders of children's work samples). Through such techniques, student progress is revealed by patterns in behaviors over time (British National Writing Project, 1987; Dixon & Stratta, 1986; Genishi & Dyson, 1984; Graves, 1983; Jaggar & Smith-Burke, 1985; Newkirk & Atwell, 1988; *Primary Language Record*, 1988). Using folders as a basis for discussion, teachers can easily involve students in the evaluation process (Burnham, 1986; Graves, 1983; *Primary Language Record*, 1988; Simmons, 1990; Wolf, 1988), discussing with them their ways of writing and their products, articulating changes in processes and products over time and across kinds of writing activities; students are thus helped to formulate concepts about

[2] The analytic scale may not actually give much more information than a holistic scale. Freedman (1981) found that all the categories except usage were highly correlated. Freedman modified Diederich's scale by combining usage with spelling and punctuation and making separate categories for sentence structure and word choice.

"good" writing, including the variability of "good" writing across situations and audiences (Gere & Stevens, 1985; Knoblauch & Brannon, 1984).

Beyond the uses of portfolios in writing classrooms, they are being piloted in a number of other educational assessment contexts, from mathematics assessments to arts assessments to teacher assessments in the form of pilot tests for certifying teachers through the planned National Board for Professional Teaching Standards. In a discussion of the uses of portfolios to assess teachers, Bird (1988) considers the implications of borrowing the portfolio metaphor from other professions (e.g., art, design, photography). Bird argues that the educational uses of portfolios are in need of definition. For other professions, including professional writing, conventions define the nature and contents of a portfolio. In education there are no such conventions, and so according to Bird, "[T]he borrowed idea of 'portfolio' must be reconstructed for its new setting" (p. 4). Bird's concerns become particularly important if we begin to consider possible large-scale uses of portfolios. A survey of the literature on writing portfolios readily reveals that most portfolio projects lack guidance on several fundamental fronts: what writing is to be collected, under what conditions, for what purposes, and evaluated in what ways. Murphy and Smith (1990) outline a set of questions that must be answered by anyone designing a portfolio project: "Who selects what goes into the portfolio?" "What goes into the portfolio?" "How much should be included?" "What might be done with the portfolios?" "Who hears about the results?" "What provisions can be made for revising the portfolio program?" (p. 2).

As the fundamental nature of the questions indicate, portfolio assessment is finding its way into practice well before the concept has been defined. Wiggins (1990) explains that people are "doing" portfolios, but the operational definitions range broadly, the purposes vary widely, and as Bird (1988) points out, the underpinnings are metaphorical more than analytic and most likely "the potential of portfolio procedures depends as much on the political, organizational and professional settings in which they are used as on anything about the procedures themselves" (p. 2). Camp (1990) lists several essential features which contain implications for the kinds of writing and thinking activities that will have to accompany portfolios and that will influence the professional setting:

> multiple samples of classroom writing, preferably collected over a sustained period of time;
>
> evidence of the processes and strategies that students use in creating at least some of those pieces of writing;
>
> evidence of the extent to which students are aware of the processes and strategies they use in writing and of their development as writers. (p. 10)

Still, the unifying theme is little more than "collecting 'real' student work," including information about students' processes and their reflections on their work.

Before turning to the potential of portfolios to inform large-scale testing, I will first illustrate the concept by showing how portfolios are being integrated into a school system. Wolf (1988, 1989a,b) writes about Arts PROPEL, a school-district portfolio project in art, music, and imaginative writing designed as a collaborative with the Pittsburgh public schools, Harvard's Project Zero, and the Educational Testing Service. Arts PROPEL aims eventually to provide "alternatives to standardized assessment" (Wolf, 1989a), but first is exploring the power of portfolios to impact teaching and learning, to change educational settings:

> Central to this work [the portfolio project] are two aims. The first is to design ways of evaluating student learning that, while providing information to teachers and school systems will also model [the student's] personal responsibility in questioning and reflecting on one's own work. The second is to find ways of capturing growth over time so that students can become informed and thoughtful assessors of their own histories as learners. (p. 36)

According to Wolf, teachers in Arts PROPEL are concerned with the following important questions underlying thoughtful pedagogy, appropriate assessment, and professionalized school settings:

> • *How do you generate samples of work which give a genuine picture of what students can do?*
>
> • *How do you create "three-dimensional" records—not just of production, but of moments when students reflect or interact with the work of other writers and artists?*
>
> • *How do you invite students into the work of assessment so that they learn life-long lessons about appraising their own work?*
>
> • *How could the reading of portfolios turn out to be a situation in which teachers have the opportunity to talk with one another about what they value in student work? About the standards they want to set; individual differences in how students develop; conflicts between conventions and inventions?* (1989b, p. 1)

Wolf is quick to point out the importance of taking such questions seriously:

> Portfolios are not MAGIC. Just because students put their work into manila folders or onto tapes, there is no guarantee that the assessment that follows is wise or helpful. The assignments could be lockstep. Students could be asked to fill out worksheets on reflection. The portfolio could end up containing a chronological sample of short answer tests. Scoring might be nothing more than individual teachers counting up assignments or taking off points for using the wrong kind of paper. (p. 1)

Currently, the Arts PROPEL portfolio data are not used for any assessment purpose beyond classroom teaching and school-level coordination of information.

Moving Toward Large-Scale Portfolio Use: In Schools, in State Testing Programs, and for National Examinations in Great Britain

How can we begin to link classroom portfolios to assessment and testing goals beyond the classroom? A start of an answer comes from a second example of portfolios in classroom use, but on a larger-scale than Arts PROPEL and with some attempts at standardization of information collected: *The Primary Language Record (PLR)*, developed in Great Britain. The *PLR* is designed to introduce systematic record-keeping about language growth, a kind of portfolio, into all elementary classrooms in the U.K. The *PLR* was written by a committee of teachers and administrators at varied levels and piloted in more than 50 schools to refine the final version. The classroom teacher collects the portfolios for three reasons: "to inform and guide other teachers who do not yet know the child; to inform the headteacher and others in positions of responsibility about the child's work; to provide parents with information and assessment of the child's progress" (1988, p. 1). The British argue that all assessment should be formative and qualitative until the end of secondary school and hence the *PLR* is designed as a qualitative assessment tool, but one that provides specific directions and even standard forms on which to collect and record children's language growth.

For the writing portion of the record, teachers are asked to "Record observations of the child's development as a writer (including stories dictated by the child) across a range of contexts" (p. 44). Teachers are directed to consider:

> —the child's pleasure and interest in writing
>
> —the range and variety of her/his writing across the curriculum
>
> —how independent and confident the child is when writing
>
> —whether the child gets involved in writing and sustains that involvement over time
>
> —the child's willingness to write collaboratively and to share and discuss her/his writing
>
> —the understanding the child has of written language conventions and the spelling system (p. 44)

Teachers are also asked to record observations about children's writing samples at least "once a term or more frequently" (p. 50).[3] The writers of the *PLR* note that "Many schools already collect examples of children's writing in folders which become cumulative records"; the method of sampling they are suggesting "draws on that practice and allows for the systematic collection and analysis of work." They claim that the *PLR* adds "a structured way of looking in depth at particular pieces of writing" (p. 50). In guiding these structured and in-depth looks at samples of student work, the *PLR* asks for the inclusion of: "1 Context and background information about the writing. . . . 2 Child's own response to the writing. . . . 3 Teacher's response . . . . 4 Development of spelling and conventions of writing. . . . 5 What this writing shows about the child's development as a writer" (pp. 51-52).

An example of a six-year-old boy's writing and the sample *PLR* entries about it make clear what the record contributes:

> One day annansi met hare and they went to a tree fooll of food annansi had tosing a little soing to get the rope and the rope did Not come dawn its self his mother dropt it dawn and he climb up it hoe towld hare not to tell but at ferst he did not tall but in a little wille he did.
>
> He towlld eliphont and the tottos and the popuqin and the caml and they saing the little soing and dawn came the rope and they all clambd on it and the rope swuing rawnd and rawnd.
>
> and they all screemd and thir screemds wock Anansi up and he shawtdid to his mother it is not Anansi but robbers cut the rope.

[3] In the U.K. the school year is divided into three terms: fall, winter, and summer.

> and she cut the rope and anmls fell and the elphent flatnd his fas and the totos crct his shell and the caml brocka bon in his humpe and pocupin brock all his pricls. (p. 51)

The teacher writes first about the context and background of the story:

> M. wrote this retelling after listening to the story on a story tape several times. Probably particularly interested in it because of the Caribbean stories told by storytellers who visited recently. Wrote the complete book in one go—took a whole morning. First draft. (p. 51)

The child's response:

> Very pleased with it. He has talked a lot about the story since listening to the tape. (p. 51)

The teacher's response:

> I was delighted. It's a very faithful retelling, revealing much detail and language. It's also a lengthy narrative for him to have coped with alone. (p. 51)

About the student's developing control of spelling and conventions, the teacher continues:

> He has made excellent attempts at several unfamiliar words which he has only heard, not read, before. Apart from vowels in the middle of words he is getting close to standard spelling. (p. 51)

Finally, about his general development, the teacher concludes:

> It is the longest thing he's done and the best in technical terms. He is happy with retelling and likes to have this support for his writing, but it would be nice to see him branching out with a story that is not a retelling soon. (p. 51)

Basically, what the *PLR* provides is a guide to the teacher for commenting on student's work and for keeping a running record that can be accessed by others. The *PLR*, although more specific than any other writing on classroom portfolios, remains relatively vague. For example, the following is only guidance for the teacher response category of the *PLR*:

> Is the *content* interesting? What about the *kind of writing*—is the child using this form confidently? And finally, how does this piece strike you as a reader—what is your reaction to it? (p. 52)

The *PLR* also does not suggest how qualitative comments could be systematically aggregated to provide information about anything other than individual development. Certainly, the push to create classroom portfolios has great potential for improving teaching and learning. And the records being kept might become useful to large-scale testers, if we could begin to figure out some sensible ways not just to collect but also to make use of the data for determining how well students can write, how effective our curriculum is.

In the U.S. we are mostly at the stage of experimenting with putting portfolio evaluation systems in place at the classroom and school level in sensible ways, without worrying too much about their wider uses. However, the hope is, as Wolf writes, that portfolios will someday replace more traditional forms of large-scale assessment. Toward this end, a number of states have begun to support portfolio development work in school settings, basically allowing creative teachers and administrators to "mess around" with portfolios, tailoring them to local contexts, seeing what happens. For example, California has funded several school-site efforts (see Murphy & Smith, 1990). In Alaska three districts are being funded to create integrated language arts portfolios: a high school in Fairbanks is having students put together portfolios to be judged as part of a graduation/exit test; a first-grade classroom in Juneau is using portfolios instead of report cards and is also using them to determine gains for Chapter 1 programs and for decisions about promotion to grade 2; and two elementary school-wide projects are being put in place in Anchorage. [4]

The state of Vermont is perhaps farther along than most others in conceptualizing a state-wide portfolio assessment program. The Vermont experience is showing how assessment goals and classroom reform can be coupled, and mutually supported; however, for now the coupling is more like an engagement than a marriage since the plan is still only a plan. A draft of the plan, *Vermont Writing Assessment: THE PORTFOLIO* (1989), announces:

> We have devised a plan for a state-wide writing assessment that we think is humane and that reinforces sound teaching practices. . . . As a community of learners, we want to discover, enhance and examine good writing in Vermont. As we design an assessment program, we hope to combine local common sense with the larger world of ideas . . . and people. . . . We believe that guiding students as writers is the responsibility of every teacher and administrator in the school and that members of the public have a right to know the results of our efforts. (p. 1)

Vermont plans to assess all students in grades four and eleven. The plan has three parts. First, students will write one piece to an assigned and timed prompt which will be holistically scored. Second, with the help of their classroom teacher, students will select and submit a "best piece" from their classroom writing portfolio. This piece will be scored by the same teachers who evaluate the prompted sample. Finally, state evaluation teams will visit all schools "to review a sample of fourth and eleventh grade portfolios" (p. 2). At this time the "teams will look at the range of content, the depth of revision and the student's willingness to take a risk" (p. 2). The idea is that "scores from the prompted sample and the best piece will indicate each student's writing abilities; portfolios will give a picture of the school's writing program" (p. 2).

For the classroom portfolios the Vermont draft plan advises that students keep "all drafts of any piece the student wants included" (p. 3). The plan also advises schools to buy or clear storage cabinets. The idea is that students will keep this full "current-year folder" which will then be transferred to a permanent folder which will include a selected collection of the students' work from grades kindergarten through grade 12. The current year folder will contain a cover sheet much like that just described in the *PLR*. It will have space for teacher comments, instructions and goals for the students, and the state evaluation team's official comments, along with a grid/checklist for documenting the process of

[4] Other states implementing or experimenting with portfolio assessment include: Alaska, Arizona, California, Connecticut, Maryland, New Mexico, Oregon, Texas, and Rhode Island. States that have expressed interest but that do not yet have formal committees include: Arkansas, Nebraska, and Utah. This information was compiled through 1990 telephone interviews with officials at each state department of education by Pamela Aschbacher of the Center for the Study of Evaluation at UCLA.

producing the portfolio work. For inclusion in the portfolio, the state team will likely recommend a minimum set of pieces of varied types, *either* something expressive, imaginative, informative, persuasive, and formulaic (to fulfill social obligations) or alternatively a letter explaining the choices of work in the portfolio, *or* a piece about the process of composition, a piece of imaginative writing, a piece for any non-English curriculum area, and a personal written response to a book, current issue or the like.

The plan for the teachers' evaluating of the portfolio follows: "To assess student portfolios, we propose asking teacher-evaluators to answer a set of questions, using a format that allows for informal and formal portfolio reviews" (p. 13). The questions include both a scale, with a numerical score and a place for qualitative comments. For example, the first of the 14 scaled questions is:

CHECK BOXES (INFORMAL)	GRADUATED TERMS (INFORMAL)	(FORMAL) HOLISTIC SCORE
□ 1.	<u>DOES WRITING REFLECT A SENSE OF AUTHENTIC VOICE?</u>	2 3 4 5 6 7 8
	□ Somewhat □ Consistently □ Extensively	

Other questions ask about audience awareness, logical sequence, syntax, and spelling as well as about the process the student used to produce the pieces and the folder, and about the coherence of the folder as a whole. The qualitative comment section is like the *PLR* only less elaborate, with only a space for general observations and another for recommendations.

The Vermont plan is comprehensive and involves provision for teacher in-service in the collection and evaluation of student portfolios as well as for a state-wide evaluation that takes into account student writing produced under both natural and testing conditions. In addition, through the site teams, Vermont has a plan for evaluating programs at the school site level. Although still in the planning stages, Vermont seems to be leading the way in connecting teacher in-service and assessment with the large-scale evaluation of writing programs and testing of writing. This coordinated plan promises to provide information about the development of individual students, about school programs, and about writing achievement in the state.

As a final example of the large-scale use of portfolios, I want to turn to the national examination that determines whether or not British students at age 16+,,the end of U.S. tenth-grade equivalent, will graduate from secondary school and receive the equivalent of a U.S. high school diploma. This British examination is called the General Certificate of Secondary Education (GCSE).[5] If students receive high scores on the GCSE, they may go into a two-year course, the General Certificate of Education at Advanced Level, known as A levels. The A level courses qualify students for entry to universities and other forms of higher education. Also, some employers demand A levels. Over 60% of U.K. students do not take A levels but instead leave school at 16+, after taking the GCSE examination. The GCSE serves a major gatekeeping function in Great Britain.

[5] The GCSE has replaced the system by which more able students, the top 20-25%, were entered for the General Certificate of Education Ordinary level (O level) and others took the Certificate of Secondary Education (CSE).

For the GCSE in language and literature, schools choose between *either* a timed examination at the end of the two years *plus* a folder of coursework (portfolios) *or* simply a folder of coursework. The important point is that the GCSE now contains coursework and is in large part or is completely a national, large-scale examination, based on portfolios of students' coursework. In the case of the English and language examinations, the coursework is writing. The specifications for the GCSE differ slightly according to five different examining boards in England and Wales. For the GCSE examination schools have a choice of affiliating with any one of the five boards, each with a different examination syllabus, i.e., format and organization for the examination as well as the course of study.

For the coursework only option, students must complete 20 pieces of writing, ten of these for the English language examination and ten for the literature examination, the two examinations being separately assessed. The writing in the folder must be in a variety of functions, for a variety of purposes, and for different audiences (e.g., report, description, argument, and persuasion, narrative fiction, poems, response to texts), assembled over a two-year period (usually with the same teacher for both years of the examination course), on which the students' grades are totally based. Of the ten pieces for each examination, the student and teacher choose the five best pieces which cover the assessment objectives for each examination. These are the pieces which are finally evaluated.

For this coursework only option, the assessment of the writing in the coursework folder is made by the student's teacher, by a committee of teachers in the school, and is checked and standardized nationally. The national standard-setting for portfolio marking is done somewhat differently by the different examining boards, but the general plans are quite similar. A booklet produced by the NEA reports that representatives from each school who are teachers and are involved in the national standard setting meet twice a year for trial marking sessions where they receive photocopies of scripts or portfolios entered by four students the previous year. The portfolios do not have grades, so the teachers decide the grade they would give if the candidate was their student. The teachers submit their grades at a school meeting where the portfolios are discussed and a school grade agreed upon. Representatives from each school attend a consortium trial marking meeting where portfolios and grades are discussed again. A member of the NEA's National Review Board attends this meeting and explains the grades the Board has given. After this training period a committee of teachers in the school agrees on grades for the coursework folders from that school (at least two teachers from the committee have to agree on the grade), and then the folders are sent to a review panel where the reviewers evaluate a sample from each school. If the National Board consistently disagrees with the evaluations from a school, all portfolios from that school are regraded. The final grade for the student is then sent back to the school.

The important point is that the student's examination grades for language and then for literature are based on an evaluation for the set of pieces in that area in the folder. The portfolio evaluation consists of a grade given for a group of pieces and is not derived from an average of grades on individual pieces. All assessors, including the National Review Panel, are practicing teachers.

The GCSE is elaborate and standardized, both in the plan for marking the folders and in the plan for collecting the work that goes in them. The GCSE also shows the crucial role the teacher plays in the student's success on a portfolio evaluation. Teachers always play this role, of course, but portfolios place the responsibility inequivocably and directly in the teacher's lap.

The *PLR*, the Vermont plan, and the GCSE illustrate several ways that portfolio assessment can be used, with the assessment designs appropriately varied according to the functions they fulfill. Although the models of well-conceived, large-scale portfolio programs are few, they are certainly beginning to emerge, and they are marked by their thoughtful approach to students and to the evaluation of their work.

Conclusions

In the assessment of writing, the concept of the portfolio seems particularly appealing because writers, like artists, can collect representative samples of their work that provide a sense of the range and quality of what they can do (Anson, Bridwell-Bowles, & Brown, 1988; Burnham, 1986; Camp, 1985a,b, 1990; Elbow & Belanoff, 1986a,b; Fowles & Gentile, 1989; Lucas, 1988a,b; Murphy & Smith, 1990; Simmons, 1990; Stiggins, 1988; Wolf, 1988, 1989a). Portfolios can be collected as part of an ongoing instructional program and get around the problem of one-shot evaluation procedures (Anson et al., 1988; Belanoff, 1985; Burnham, 1986; Calfee & Sutter-Baldwin, 1987; Calfee & Hiebert, 1988; Camp, 1985a,b; Camp & Belanoff, 1987; Elbow, 1986; Elbow & Belanoff, 1986a,b; Fowles & Gentile, 1989; Lucas, 1988a,b; Murphy & Smith, 1990; Simmons, 1990; Valencia, McGinley, & Pearson, 1990; Wolf, 1988). Providing direction for large-scale portfolio efforts that could inform and be informed by classroom efforts is particularly important, since testing programs often exert powerful influences over the nature of instruction in writing and reflect "what counts" as literacy (Calfee & Hiebert, 1988; Cooper, 1981a; Cooper & Murphy, in progress; Cooper & Odell, 1977; Diederich, 1974; Loofbourrow, 1990; Mellon, 1975; Myers, 1980; Resnick & Resnick, 1977, 1990). There is an important role for teacher-driven and classroom-based assessment in our plans for educational reform.

But I want to end with a word of warning. Currently, in the U.S. the National Assessment is experimenting with the collection of information from writing portfolios. Preliminary results are showing that when a random group of teachers are just asked to submit student work, called portfolios, without the accompanying staff development and professional activities outlined in most of the programs I have described, the writing that they submit is rather dismal. As the careful work of the Pittsburgh Arts PROPEL project shows, just collecting and evaluating portfolios will solve neither our assessment problems, nor our need to create a professional climate in our schools. By coupling assessment and instruction in increasingly sophisticated ways, we may be able to make a real difference in education in this country. What I have offered here is an overview of writing assessment and some examples of programs that might stimulate us to think about new directions.

References

Anson, C., Bridwell-Bowles, L., & Brown, R. L., Jr. (1988, April). *Portfolio assessment across the curriculum: Early conflicts.* Three papers presented at the National Testing Network in Writing, Minneapolis, MN. Summarized in *Notes from the National Testing Network in Writing, 8*, 6-7. New York: The City University of New York, Office of Academic Affairs, Instructional Resource Center.

Baker, E. (1989). Mandated tests: Educational reform or quality indicator? In B. R. Gifford (Ed.), *Future assessments: Changing views of aptitude, achievement, and instruction.* Boston, MA: Kluwer Academic Publishers.

Belanoff, P. (1985, November). In S. Murphy (Recorder), Models of portfolio assessment. In K. L. Greenberg & V. B. Slaughter (Eds.), *Notes from the National Testing Network in Writing* (pp. 2 & 7). New York: The City University of New York, Instructional Resource Center.

Bird, T. (1988). The schoolteacher's portfolio: An essay on possibilities. In J. Millman & L. Darling-Hammond (Eds.), *Handbook of teacher evaluation: Elementary and secondary personnel* (2nd ed.). Newbury Park, CA: Sage.

British National Writing Project. (1987). *Ways of looking at children's writing: The National Writing Project response to the task group on assessment and testing* (Occasional Paper No. 8). London: School Curriculum Development Committee Publications.

Brown, R. (1986). A personal statement on writing assessment and education policy. In K. Greenberg, H. Weiner, & R. Donovan (Eds.), *Writing assessment: Issues and strategies.* New York: Longman.

Burnham, C. (1986). Portfolio evaluation: Room to breathe and grow. In C. Bridges (Ed.), *Training the teacher of college composition.* Urbana, IL: National Council of Teachers of English.

Burstein, Baker, E., Aschbacher, P., & Keesling, J. (1985). *Using state test data for national indicators of education quality: A feasibility study.* Final report, NIE grant G-83-001). Los Angeles: Center for the Study of Evaluation.

Calfee, R., & Hiebert, E. (1988). The teacher's role in using assessment to improve learning. In E. Freeman (Ed.), *Assessment in the service of learning: Proceedings of the 1987 Educational Testing Service Invitational Conference.* Princeton, NJ: Educational Testing Service.

Calfee, R. (1987). The school as a context for the assessment of literacy. *The Reading Teacher, 40* (8), 438-443.

California State Department of Education. (1989). *Writing achievement of California eighth graders: A first look.* Sacramento, CA: California Assessment Program.

California State Department of Education. (1990). *Writing assessment handbook: Grade eight.* Sacramento, CA: California Assessment Program.

Calkins, A. (1990). Personal Correspondence.

Camp, R. (1985a, November). In S. Murphy (Recorder), Models of portfolio assessment. In K. L. Greenberg & V. B. Slaughter (Eds.), *Notes from the National Testing Network in Writing* (pp. 2 & 7). New York: The City University cf New York, Instructional Resource Center.

Camp, R. (1985b). The writing folder in post-secondary assessment. In P. J. A. Evans (Ed.), *Directions and misdirections in English evaluation* (pp. 91-99). Ottawa, Canada: The Canadian Council of Teachers of English.

Camp, R. (1990). Thinking together about portfolios. *The Quarterly of the National Writing Project and the Center for the Study of Writing*, (12) 2, 8-14, 27.

Camp, R., & Belanoff, P. (1987). Portfolios as proficiency tests. *Notes from the National Testing Network in Writing*, 7 (8).

Charney, D. (1984). The validity of using holistic scoring to evaluate writing: A critical overview. *Research in the Teaching of English, 18* (1), 65-81.

Conlan, G. (1986). "Objective" measures of writing ability. In K. L. Greenberg, & V. B. Slaughter (Eds.), *Notes from the National Testing Network in Writing* (p. 2 & 7). New York: The City University of New York, Instructional Resource Center.

Cooper, C. R. (1977). Holistic Evaluation of Writing. In C. Cooper & L. Odell (Eds.), *Evaluating Writing*. Urbana, IL: National Council of Teachers of English.

Cooper, C. R. (1981a). Competency testing: Issues and overview. In C. R. Cooper (Ed.), *The nature and measurement of comptency in English.* Urbana, IL: National Council of Teachers of English.

Cooper, C. R., & Murphy, S. (in progress). *A report on the CAP Writing Assessment and its influences on the classsroom.*

Cooper, C. R., & Odell, L. (Eds.). (1977). *Evaluating writing: Describing, measuring, judging.* Urbana, IL: National Council of Teachers of English.

Davis, B., Scriven, M., & Thomas, S. (1987). *The evaluation of composition instruction* (2nd ed.). New York: Teachers College Press.

Diederich, P. (1974). *Measuring growth in English.* Urbana, IL: National Council of Teachers of English.

Diederich, P. B., French, J. W., & Carlton, S. T. (1961). *Factors in judgments of writing ability* (Research Bulletin No. RB-61-15). Princeton, NJ: Educational Testing Service.

Dixon, J., & Stratta, L. (1986). *Writing narrative—and beyond.* Upper Montclair, NJ: Boynton/Cook.

Dyson, A. H., & Freedman, S. W. (in press). Writing. In J. Jensen, J. Flood, D. Lapp, & J. Squire (Eds.), *Handbook of research on teaching the English language arts.* New York: Macmillan Publishing Co.

Elbow, P. (1986). Portfolio assessment as an alternative in proficiency testing. *Notes from the National Testing Network in Writing, 6, 3,* and *12.*

Elbow, P., & Belanoff, P. (1986a). Portfolios as a substitute for proficiency examinations. *College Composition and Communication, 37* (3), 336-337.

Elbow, P., & Belanoff, P. (1986b). Using portfolios to judge writing proficiency at SUNY Stony Brook. In P. Connolly & T. Vilardi (Eds.), *New directions in college writing programs*. New York: Modern Language Association.

Faigley, L., Cherry, R. D., Jolliffe, D. A., & Skinner, A. M. (1985). *Assessing writers' knowledge and processes of composing*. Norwood, NJ: Ablex.

Fowles, M., & Gentile, C. (1989). *The fourth report on the New York City junior high school writing and learning project: Evaluation of the students' writing and learning portfolios (March 1989-June 1989)*. Princeton, NJ: Educational Testing Service.

Freedman, S. W. (1979). How characteristics of student essays influence teachers' evaluations. *Journal of Educational Psychology, 71*, 328-338.

Freedman, S. W. (1981). Influences on evaluators of expository essays: Beyond the text. *Research in the Teaching of English, 15*(3), 245-255.

Freedman, S. W. (1987). *Response to student writing*. Research report 23. Urbana, IL: National Council of Teachers of English.

Genishi, C., & Dyson, A. H. (1984). *Language assessment in the early years*. Norwood, NJ: Ablex.

Gere, A. R., & Stevens, R. (1985). The language of writing groups: How oral response shapes revision. In S. W. Freedman (Ed.), *The acquisition of written language: Response and revision* (pp. 85-105). Norwood, NJ: Ablex.

Godshalk, F. I., Swineford, F., & Coffman, W. E. (1966). *The measurement of writing ability* (Research Monograph No. 6.). New York: College Entrance Examination Board.

Gorman, T., Purves, A., & Degenhart, R. (1988). *The IEA study of written composition I: The international writing tasks and scoring scales*. Oxford: Pergamon Press.

Gubb, J., Gorman, T., & Price, E. (1987). *The study of written composition in England and Wales*. Windsor, England: NFER-NELSON Publishing Company Ltd.

Graves, D. H. (1983). *Writing: Teachers and children at work*. Portsmouth, NH: Heinemann Educational Books.

Huddleston, E. (1954). Measurement of writing ability at the college-entrance level: Objective vs. subjective testing techniques. *Journal of Experimental Psychology, 22*, 165-213.

Jaggar, A., & Smith-Burke, T. (1985). *Observing the language learner*. Urbana, IL: National Council of Teachers of English.

Knoblauch, C., & Brannon, L. (1984). *Rhetorical traditions and the teaching of writing*. Upper Montclair, NJ: Boynton-Cook.

Lloyd-Jones, R. (1977). Primary trait scoring. In C. Cooper and L. Odell (Eds.), *Evaluating writing*. Urbana, IL: National Council of Teachers of English.

Loofbourrow, P. (1990). *Composition in the context of CAP: A case study of the influence of the California Assessment Program on composition in one junior high school*. Unpublished doctoral dissertation, University of California, Berkeley.

Lucas, C. Keech. (1988a). Recontextualizing literacy assessment. *The Quarterly of the National Writing Project and the Center for the Study of Writing, 10* (2), 4-10.

Lucas, C. Keech. (1988b). Toward ecological evaluation. *The Quarterly of the National Writing Project and the Center for the Study of Writing, 10* (1), 1-3, 12-17.

Mellon, J. C. (1975). *National assessment and the teaching of writing: Results of the first National Assessment of Educational Progress in writing*. Urbana, IL: National Council of Teachers of English.

Meyers, A., McConville, C., & Coffman, W. (1966). Simple structure in the grading of essay tests. *Educational and Psychological Measurement, 26*, 41-54.

Murphy, S., & Smith, M. A. (1990). Talking about portfolios. *The Quarterly of the National Writing Project and the Center for the Study of Writing*, (12) 2, 1-3, 24-27.

Myers, M. (1980). *A procedure for writing assessment and holistic scoring*. Urbana, IL: National Council of Teachers of English.

Newkirk, T., & Atwell, N. (1988). *Understanding writing: Ways of observing, learning and teaching* (2nd ed.). Portsmouth, NH: Heinemann.

National Assessment of Educational Progress. (1986a). *The writing report card: Writing achievement in American schools*. Princeton NJ: Educational Testing Service.

National Assessment of Educational Progress. (1986b). *Writing: Trends across the decade, 1974-1984*. Princeton, NJ: Educational Testing Service.

National Assessment of Educational Progress. (1990a). *Learning to Write in our Nation's Schools: Instruction and achievement in 1988 at grades 4, 8, and 12*. Princeton, NJ: Educational Testing Service.

National Assessment of Educational Progress. (1990b). *The writing report card, 1984-88: Findings from the nation's report card*. Princeton, NJ: Educational Testing Service.

Nold, E. (1981). Revising. In C. H. Fredericksen & J. F. Dominic (Eds.), *Writing: The natur, development, and teaching of written communication: Vol. 2. Process, development and communication* (pp. 67-80). Hillsdale, NJ: Erlbaum.

The primary language record: Handbook for teachers. (1988). London: ILEA Centre for Language in Primary Education.

Resnick, D. P., & Resnick, L. B. (1977). The nature of literacy: An historical exploration. *Harvard Education Review, 47* (3), 370-385.

Resnick, L., & Resnick, D. (1990). Tests as standards of achievement in schools. In J. Pfleiderer (Ed.), *The uses of standardized tests in American education: Proceedings of the 1989 Educational Testing Service Invitational Conference*. Princeton, NJ: Educational Testing Service.

Silberman, A. (1989). *Growing up writing*. New York: Time Books.

Simmons, J. (1990). Portfolios as large-scale assessment. *Language Arts, 67* (3), 262-268.

Stiggins, R. J. (1988, January). Revitalizing classroom assessment: The highest instructional priority. *Phi Delta Kappan*, 363-368.

Valencia, S., McGinley, W., & Pearson, P. D. (1990). Assessing literacy in the middle school. In G. Duffy (Ed.), *Reading in the middle school* (2nd ed.). Newark, DE: International Reading Association.

Vermont writing assessment: THE PORTFOLIO. (1989). Montpelier, VT: Vermont Department of Education.

White, E. (1985). *Teaching and assessing writing*. San Francisco: Jossey-Bass Publishers.

Wiggins, G. (1990). "Standards" should mean "Qualities," not "Quantities." *Education Week.*

Witte, S. P., Cherry, R., Meyer, P., & Trachsel, M. (in press). *Holistic assessment of writing: Issues in theory and practice*. New York: Guildford Press.

Wolf, D. P. (1988). Opening up assessment. *Educational Leadership, 45* (4), 24-29.

Wolf, D. P. (1989a). Portfolio assessment: Sampling student work. *Educational Leadership, 46* (7), 4-10.

Wolf, D. P. (1989b). When the phone rings. *Portfolio: The Newsletter of Arts PROPEL, 1* (5), 1.

Author's Note

I want to thank a number of people who were generous with their time in helping me think about and gather information for this paper. Bob Calfee provided many ideas, through discussions we have had over the years about testing and school reform and more recently as we have discussed the potentials of portfolios. Discussions with Anne Haas Dyson provided many ideas as well. From the National Assessment of Educational Progress, Lynne Jenkins and Ina Mullis answered many questions and helped me gather information about NAEP beyond the published resports. Annie Calkins of the Alaska State Department of Education provided detailed information about Alaska's experiences. Mary Fowles provided information about a number of state efforts. Pam Aschbacher took the time to synthesize the information about writing assessment in particular from her state department survey. Beth Brenneman provided information about the California Assessment Program. Leo Ruth and Shirley Brice Heath suggested readings. Finally, I thank Mary K. Healy for her suggestions for the manuscript.

NATIONAL ADVISORY PANEL
The Center for the Study of Writing

Chair
Fred Hechinger
The New York Times Foundation

Alonzo Crim
Professor of Urban Educational Leadership
Georgia State University, Atlanta, GA

Sister Regina Noel Dunn
Teacher
Villa Maria Academy, Malvern, PA

Marcia Farr
Associate Professor of English
University of Illinois, Chicago, IL

Abraham Glassman
Chairman
Connecticut State Board of Education

Bill Honig
California Superintendent
of Public Instruction

The Honorable Gary K. Hart
California State Senator

Sibyl Jacobson
Executive Director
Metropolitan Life Foundation

John Maxwell
Exeutive Director
National Council of Teachers of English

Roy Peña
Principal
Andrews High School, El Paso, TX

Carol Tateishi
Teacher
Ross Elementary School, Kentfield, CA

Richard C. Wallace, Jr.
Pittsburgh Superintendent of Schools
and Secretary, Board of Education