EDUCATIONAL ASSESSMENT, 1(1), 27-52

Linking Large-Scale Testing and Classroom Portfolio Assessments of Student Writing

Sarah Warshauer Freedman
School of Education
University of California at Berkeley

In this article, I explore how portfolio assessment in the area of writing can provide new links between large-scale testing and classroom assessment. After analyzing varied types of large-scale testing in writing—including multiple-choice tests and tests that include writing samples produced in testing settings—I describe several portfolio programs at work: one in the Pittsburgh school district, Arts PROPEL; a large-scale, classroom-centered portfolio effort for elementary students in England, *The Primary Language Record* (1988); a state-level portfolio assessment from Vermont for Grades 4 and 11; and a large-scale national examination for completion of secondary school in Great Britain, the General Certificate of Secondary Education. These examples show how portfolio assessment could link large-scale testing and instructional assessment. However, the examples also show that classroom-based portfolios were more likely to be successful and lasting than portfolios used for large-scale testing. I argue that if testing and assessment in writing are to be linked, portfolios will need to gain a stronger foothold in large-scale testing, something that will happen only when testers, teachers, and policymakers begin to work together reciprocally.

Somehow the teaching of English has been wrenched out of the Age of Aquarius 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.

Requests for reprints should be sent to Sarah W. Freedman, School of Education, University of California, Berkeley, CA 94720.

> 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-squares to measure what we've done without ever understanding what we were trying to do. (Hogan, 1974, p. iii)

These remarks pointing out gaps between teachers of writing and the testing and measurement community seem timely today, but they were written by Paul Hogan, then executive director of the National Council of Teachers of English, for the foreword to Paul Diederich's (1974) classic book, *Measuring Growth in English.* In this article, I explore the nature of these gaps. The focus is on two currently distinct kinds of writing evaluation—large-scale testing at the national, state, district, and sometimes school levels, which is the natural domain of the educational accountants, and classroom assessment by teachers observing 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.

In this article, the term *testing* refers to large-scale standardized evaluation, and *assessment* refers to the evaluative judgments of the classroom teacher. Calfee (1987) described 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' " (p. 738), whereas 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 & Drum, 1979)" (p. 738). In writing, as in most academic areas, 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, their teaching, and their students' learning.

Before presenting ideas for linking large-scale testing and classroom assessment, I review the form of most large-scale writing tests and discuss their limitations. Then I describe portfolio assessment, an important innovation in classroom writing evaluation that is filtering up in some cases to university-wide assessment programs, to state-wide testing programs across grade levels, and 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 give several examples of portfolio programs at work, examples that point toward possible future directions for writing assessment and instruction in this country: a large urban school district's use of portfolios in Arts PROPEL; a large-scale, classroom-centered portfolio effort for elementary students in England, *The Primary Language Record* (*PLR;* 1988) a state-level portfolio assessment from Vermont for Grades 4 and 11; and a large-scale national examination for completion of secondary school in Great Britain, the General Certificate of Secondary Education (GCSE).

LARGE-SCALE TESTING

Historically, the large-scale testing of writing has developed to: (a) determine whether students have mastered writing at some level (e.g., the NAEP); (b) evaluate writing programs in the school, the district, or in some cases the classroom (e.g., the California Assessment Program); (c) place students in programs or classes (e.g., many college placement examinations given to entering freshmen, usually administered on local campuses but sometimes, as in California, in the form of state-wide tests); and (d) decide the fate of individuals with respect to admissions, promotion, or graduation (gatekeeping; e.g., the SAT, high school and college graduation tests, and 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 difficult problems: how to evaluate student writing reliably, cost effectively, and fairly. One highly criticized but commonly used way of evaluating 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 were in widespread use as late as 1984. At that time, 19 state departments of education measured the writing of students in kindergarten through 12th grade indirectly; only 13 state departments had direct measures, and 18 had no measures at all (Baker, 1989; Burstein, Baker, Aschbacher, & Keesling, 1985). The appeal of indirect measures of writing was obvious; they were quick to administer and cheap to score. The problems were obvious too; indirect measures were poor predictors of how well the test-taker actually writes. As Conlan (1986) emphasized:

> 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 L. B. Resnick and D. P. Resnick (1990) that "you get what you assess" (p. 66), multiple-choice writing tests are bound to have negative effects on instruction because teaching to the test would not include asking students to write. There is also evidence that "black and Hispanic background students perform at about one standard deviation below the mean on standardized tests of intelligence, aptitude, and achievement (Samuda, 1975; Padilla, 1979; Olmedo, 1981; Green, 1981)" (O'Connor, 1989, p. 129). The problems for these students, who are also nonnative speakers of English, was exacerbated on multiple-choice tests of writing because what the exams test is often knowledge of the rules of standard grammar, something a native speaker will inevitably control better than a nonnative speaker, regardless of his or her abilities as a writer.

From 1890 to 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 et al. conducted a study at the Educational Testing Service (ETS) in which "sixty distinguished readers in six occupational fields" read 300 papers written by college freshmen (cited 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 the ETS researchers understand why readers disagreed.

During the 1960s, the 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 kind of 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 determined 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, Cherry, Jolliffe, & Skinner, 1985; Myers, 1980; White, 1985). A 1991 survey by Pelavin Associates shows that 31 states had implemented some sort of direct assessment of writing performances and another 7 were developing direct assessments. These figures represent a dramatic shift from 1984, when only 13 states used direct measures.

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 are dated now. With large-scale, direct writing assessments 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. First, many tensions center around the nature of test writing itself. Although controlled and written under unaided conditions, as Diederich pointed out, such writing has little function for students other than for them to be evaluated. Second, 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, 1988b; Simmons, 1990; Witte, Cherry, Meyer, & Trachsel, 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 "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" (1986a, p. 6) across the years. The NAEP gathers informative, persuasive, and imaginative writing samples from students at the three grade levels. For 8th and 12th 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 4th graders, the

blocks last only 10 min (NAEP, 1990a). This means that 4th graders have had between 5 and 10 min each to produce up to four pieces of writing during a 30-min test; 8th and 12th graders have had between 7½ and 15 min each to produce up to four pieces during a 45-min test (NAEP, 1990a).

Writing researchers and educators have critiqued the NAEP, 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 authors of the NAEP (1990b) caution that:

> 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. (p. 7)

Based on the NAEP (1986a) writing data, how confident can we be of the following claim made in *The Writing Report Card, 1984–88:* "A major conclusion to draw from this assessment is that students at all grade levels are deficient in higher-order thinking skills" (p. 11)? What of their higher order thinking skills can students possibly reveal in 15 min when writing on an assigned topic that they have never seen?

In contrast to most testing conditions and consistent with common 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 that 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, 1991; Freedman, 1987). Most tightly timed, test-type writing goes against current pedagogical trends. What Mellon (1975) pointed out about the NAEP 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)

The NAEP is beginning to respond to criticisms about the time for the testing. In 1988, the NAEP gave a subsample of the students twice as much time on one informative, persuasive, and imaginative topic at each grade level (20 min for Grade 4 and 30 min for Grades 8 and 12; NAEP, 1990a). The results show that with increased time all students scored significantly better on the narrative tasks, and 4th and 12th 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 Black or Hispanic students, widening the gaps between these groups in the assessment results.

For the 1992 assessment, the NAEP was set up to provide more time across the board:

> As a result of both the findings from this study [of the effects of increased time for writing] 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 min or even 50 min for writing on a given topic does not resolve the basic discrepancy between what the profession argues should be happening in classrooms and what actually happens in this testing setting. Furthermore, the findings about Blacks and Hispanics raise a new set of questions about equity and testing.

To respond to these issues, in 1990 the NAEP collected portfolios of the writing of fourth and eighth graders produced as a natural part of writing instruction. The NAEP's goal was to explore alternatives to the present assessment (Gentile, 1992). For this pilot study, teachers were only given

several days notice that they were to submit portfolios, and those that submitted mostly taught in advantaged urban communities. The resulting student sample, then, was not nationally representative. The findings show low levels of correlation between the portfolios that were collected and students' performance on the NAEP exercise, especially for the eighth graders, 45% of whom received different scores on the NAEP exercise and their school-based writing. Portfolios also have been collected in 1992, allowing NAEP to continue to gather supplementary information for the main assessment and study options for future directions.

It is important to remember that, as the assessment changes, the only way to collect data about trends across time is to keep some tasks parallel. Thus, for the time being, 15-min samples are still used for the trend studies, and conclusions about trends are based only on samples produced in this 15-min time span.

Another major point of tension in the NAEP centers around the issue of scoring. To obtain more information than a single holistic score and to define clearly the features of writing being judged, in the mid-1970s the NAEP developed an additional scoring system, "the Primary Trait Scoring method" (Lloyd-Jones, 1977, p. 33). Although 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 samples on a particular topic. Also, which trait is primary and whether one aspect of writing should be labeled primary is subject to debate.

The dilemmas come across clearly through an analysis of Lloyd-Jones's (1977) example of a primary trait scoring rubric. Lloyd-Jones explained 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 writing sample 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. According to the primary trait rubric, the writing receives a score of 0 if the writer gives no response or a fragmented response; it receives a 1 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, perhaps showing how a woman has many places in the home and out? This student would receive a score of 1 but may write a substantially better essay than a student who receives a score of 2, 3, or 4 for taking a side and providing one or more reasons. In another scenario, a student who gives one elaborated reason for a score of 3 can write a far better essay than the student who gives four or more unelaborated reasons and receives a 4. Although the NAEP still advocates primary trait scoring, the rubrics have become less specific over the years and therefore less controversial.

Besides these issues of judging elaboration particular to this scoring rubric, the primary trait score measures only 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. In a study comparing holistic and primary trait scoring, the NAEP (1986a) found that primary trait scoring does not correlate particularly well with holistic quality judgments; correlations ranged from .38 to .66 depending on the topic (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. With the more general primary trait rubrics the NAEP now uses, the effect is that the primary trait scores factor out mechanics and essentially give a holistic score on all aspects of the writing except mechanics.

Whereas the NAEP (1986b, 1990b) uses a holistic score, a primary trait score, and a mechanics score for its trend reports, the NAEP (1986b, 1990a) uses only primary trait scoring for the reports on the status of writing for a given year. In the latest status report, the NAEP (1990a) explained, "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 the NAEP. States with direct writing assessments face the same challenges as the NAEP, and several states are meeting the challenges in interesting ways. For example, in 1988 in an effort to increase accountability, the Alaska State School Board mandated the Iowa Test of Basic Skills for Grades 4, 6, and 8. The Iowa test 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 were well organized through the Alaska Writing Consortium, an affiliate of the National Writing Project, and had 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 may provide. The state funded an experiment at the 10th-grade level; in 1989, 12 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. The analytic scale differs from primary trait because the categories are generic to good writing and thus are 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 studies of written language (Gorman, Purves, & Degenhart, 1988; Gubb, Gorman, & Price, 1987). Ironically, although the analytic scale provides a variety of scores, it may not actually give much more information than a holistic scale. Freedman (1981) found that all the categories except usage were highly correlated. In Freedman's study spelling and punctuation were part of usage, but sentence structure and word choice were considered separate categories.

For the Alaska test, teachers wanted to maintain some control over the testing conditions while allowing students more natural and comfortable writing conditions than are usual for large-scale, formal assessments. Thus, students were given a common prompt but were allowed two 50-min 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. In 1990, 22 districts out of Alaska's 54 districts volunteered to participate in the experiment.

NEW DIRECTIONS: WRITING PORTFOLIOS

Besides providing information about the status quo, the 1991 Pelavin Associates survey showed an emerging trend. A small number of states had

moved to or were developing a new kind of assessment, portfolio approaches. In writing, portfolios included a variety of types of writing created in the classroom across a span of time, often as long as 1 year. At the time of the 1991 survey, the only state with a portfolio assessment in place for writing was Vermont. Three other states reported being engaged in developing portfolio assessments: California, Delaware, and Rhode Island. Since then there is evidence that these state-level shifts are continuing. For example, Kentucky now uses portfolios. Alaska too has been experimenting with portfolio assessment. Aschbacher (1991) found that Arizona, Connecticut, Maryland, New Mexico, Oregon, and Texas were experimenting with portfolios and that several states—Arkansas, Nebraska, North Carolina, Utah, Wisconsin, and Wyoming—expressed interest but did not yet have formal committees. Calfee and Perfumo (1992) found that Indiana and Hawaii also are exploring portfolio options.

Calfee and Perfumo (in press) pointed out that, in the area of writing, portfolio assessment often originates in individual classrooms and sometimes schools and is part of a revolution stimulated by teachers. These shifts in large-scale testing, as was the case with the 1990 pilot study for the 1992 NAEP, get their impetus from classroom reform. This phenomenon is notable because it turns the traditional links between large-scale testing and classroom assessment on their heads. Instead of large-scale testing driving instruction, changes in instruction and classroom assessment are beginning to drive large-scale testing. In this way, the portfolio movement provides a new kind of potential link between large-scale testing and classroom assessment and teaching. It honors bottom-up reforms. Especially when teachers are involved in the design of large-scale portfolio testing, this movement has the potential to stimulate professional dialogues across varied levels of the system.

In the classroom, 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 record keeping (e.g., anecdotal records and 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; *PLR*, 1988). Using folders as a basis for discussion, teachers can involve students in the evaluation process (Burnham, 1986; Graves, 1983; *PLR*, 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 ideally are helped to formulate concepts about 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 and for large-scale testing of writing, portfolios are being piloted in a number of other contexts, from mathematics assessments to arts assessments to teacher certification through the National Board for Professional Teaching Standards. In a discussion of the uses of portfolios to assess teachers, Bird (1988) considered the implications of borrowing the portfolio metaphor from other professions (e.g., art, design, and photography). He argued 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, "The borrowed idea of 'portfolio' must be reconstructed for its new setting" (p. 4). Bird's concerns become particularly important when considering possible large-scale uses of portfolios.

A survey of the literature on writing portfolios reveals that most classroom-based 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) outlined 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 indicates, portfolio assessment is finding its way into classroom practice well before the concept has been defined. As Bird (1988) pointed 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) listed several essential features that 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.

To show the potential links between portfolios used to enrich instruction and portfolios used for large-scale testing, first I provide an example of

several varied uses. Arts PROPEL in Pittsburgh shows how portfolios can be integrated into a school system to enrich instruction. The *PLR* (1988) in Great Britain shows a classroom use that could feed into testing programs; for the *PLR,* a national model for collecting portfolios was created by teachers and then spread into classrooms across the nation to standardize information gathered about students' progress in language and literacy development. Vermont's portfolio examination shows how a reciprocity can be built between classroom assessment and large-scale testing. Finally, the GCSE in Britain offers an example of a high-stakes national exam that determines graduation from secondary school; the GCSE directs the curriculum and offers classroom portfolio options for the test in English language and literature. These varied examples explore the potential for different kinds of links between classroom assessment and large-scale testing.

LOCAL PORTFOLIO USE: ARTS PROPEL

Wolf (1988, 1989a, 1989b) wrote 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 ETS. Arts PROPEL aims eventually to provide "alternatives to standardized assessment" (Wolf, 1989a, p. 35), but first it is exploring the portfolio's impact on teaching and learning and its power 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 (1989b), 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?* (p. 1)

Wolf (1989b) quickly pointed 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. However, Wolf's (1989b) warnings emphasize how dependent this method of assessment is on good teaching and sound guidance for students in compiling the portfolio in the first place.

MOVING TOWARD LARGE-SCALE PORTFOLIO USE WITH A SCHOOL-BASED EXAMPLE: THE *PLR* (1988)

This second example of portfolios in classroom use, the *PLR,* is on a larger scale than Arts PROPEL. The *PLR* begins to illustrate how one may begin to link classroom portfolios and testing goals beyond the classroom. The *PLR* was designed to introduce systematic record keeping about language growth into all elementary classrooms in Great Britain. The *PLR* was written by a committee of teachers and administrators at varied levels and was piloted in more than 50 schools to refine the final version. For the *PLR,* the classroom teacher collects 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" (p. 1). The British argue that all assessment should be formative and qualitative until the end of secondary school; hence, the *PLR* is designed as a qualitative assessment tool, but one that provides specific directions and 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" (*PLR,* 1988, 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" (*PLR,* 1988, p. 50). In Britain, the school year is divided into three terms: fall, winter, and summer. The writers of the *PLR* noted, "Many schools already collect examples of children's writing in folders which become cumulative records" (p. 50); the method of sampling they suggested "draws on that practice and allows for the systematic collection and analysis of work" (p. 50). The writers added "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 writers of the *PLR* asked 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).

The following 6-year-old boy's writing and the sample *PLR* (1988) entries make clear what the *PLR* 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. 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 wrote 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. (*PLR*, 1988, p. 51)

The child responded: "Very pleased with it. He has talked a lot about the story since listening to the tape" (p. 51).

The teacher responded: "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).

Writing about the student's developing control of spelling and conventions, the teacher continued: "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, regarding his general development, the teacher concluded:

> 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, the *PLR* (1988) provides 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 the only guidance provided for the teacher response category of the *PLR* (1988): "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. The push to create classroom portfolios has great potential for improving teaching and learning, but the records being kept can only become useful to large-scale testers when there are ways to make use of the data for determining how well students can write and how effective our curriculum is, not just to collect data.

U.S. educators continue to experiment with putting portfolio evaluation systems in place in individual classrooms and for school-wide assessments. Calfee and Perfumo (in press) showed that most educators do not worry about the systems' wider uses. However, the hope is, as Wolf (1989a) wrote, that portfolios will replace more traditional forms of large-scale testing. The standardization of the *PLR* and the thoughtful ways teachers have been involved in its development offer possibilities for thinking about linking the two.

A STATE TESTING EXAMPLE: VERMONT

Vermont, the first state to use portfolios, is farther along than any other state in conceptualizing a statewide portfolio testing program. The Vermont experience shows how testing goals and classroom reform can be coupled and mutually supported, as well as the difficulties of using portfolios on a large scale. A draft of the plan, *Vermont Writing Assessment: THE PORTFOLIO* (1989), announced:

> 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 planned to assess all students in Grades 4 and 8 beginning in 1991–1992 but is facing some obstacles and is moving more slowly. The assessment had two parts. First, students wrote one piece to an assigned prompt in 90 min, the Uniform Writing Assessment. Second, with the help of their classroom teacher, students created a portfolio consisting of a table of contents; a best piece; a letter to the reviewers explaining the choice of the best piece and the process for writing it; a poem, short story, play or personal narration; and a personal response. In addition the fourth graders wrote one prose piece, and eighth graders wrote three prose pieces (*"This is My Best,"* 1990–1991, p. 7). The Uniform Writing Assessment, the portfolio, and the best piece were each scored with an analytic scale that captured five dimensions of writing: purpose, organization, details, voice/tone, and usage/mechanics/grammar (*"This is My Best,"* 1990–1991, p. 3).

The Vermont plan is comprehensive and involves provision for teacher training for the collection and evaluation of student portfolios (Hewitt, 1991), as well as for a statewide evaluation that takes into account student writing produced under both natural and testing conditions. The ultimate goal of this coordinated plan is to provide information about the development of individual students, about school programs, and about writing achievement in the state.

Vermont has had to move more slowly than originally planned. The headline in an *Education Week* article by Robert Rothman appearing on May 20, 1992 read, "Vt Forced To Delay Goal of Expanding Assessment System." Although Vermont teachers remain enthusiastic about portfolio

testing and participation is growing, in 1991 and 1992, not all schools managed to complete portfolios for all their students. For this reason, test results were not reported at the school level. In addition, Vermont depends on unpaid teachers to score the portfolios.

In Vermont, the difficulties have proven greater in mathematics than in writing. The success of the portfolio depends on strong classroom support. In writing, classroom reforms provided the impetus for moving to portfolios at the state level. However, in mathematics, the tests were designed to effect changes in classrooms. Not surprising, there has been slower progress in math. Monica Nelson, Burlington, Vermont's director of curriculum and staff development, explained

> that the assessment program demanded changes in math curriculum and instruction along the lines proposed by the National Council of Teachers of Mathematics, and that few schools had moved in that direction. By contrast, she noted, writing instruction had been revised a number of years ago. (Rothman, 1992, p. 21)

A NATIONAL EXAMPLE: THE GCSE IN BRITAIN

A final example of the large-scale use of portfolios is provided by the national examination that determines whether or not British students at age 16+ (equal to the end of the U.S. 10th-grade) will graduate from secondary school and receive the equivalent of a U.S. high school diploma. This British examination is called the GCSE. If students receive high scores on the GCSE, they may go into a 2-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 British 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.

For the 1987 GCSE, schools were able to choose the option of coursework (a portfolio of writing) as the only basis for evaluation. Before the coursework-only option, students were evaluated solely by their performance on a terminal examination at the end of the 2-year course. The terminal examination consisted of impromptu essay questions and writing prompts, given in a test setting. For the 1994 exam, the British government has eliminated the 100% coursework option. The 1994 language exam will consist of 20% coursework, 60% terminal examination, and 20% oral examinations. The literature exam will consist of 30% coursework and 70% terminal examination. For each terminal examina-

tion, students will write two papers. The first paper topic will be provided 2 to 3 months prior to the exam date, and the second will be assigned during the examination. The British government argued that a more traditional examination format was necessary because it feared that the 1987 changes eroded standards because teachers could help their students prepare their folders.

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). I show the syllabus from the Northern Examining Association (NEA) before and after the recent changes.

For the 1987 coursework-only option, students had to complete 20 pieces of writing—10 for the English language examination and 10 for the literature examination; the two examinations were separately assessed. For each exam, students designated 5 of the 10 pieces to be graded. The writing in the folder had to fulfill a variety of functions and be written for a variety of purposes and for different audiences (e.g., report, description, argument, persuasion, narrative fiction, poems, and response to texts), which was assembled over a 2-year period (usually with the same teacher for both years of the examination course). The students' grades were based totally on the folder's assessment.

The requirement of 20 pieces with 10 assessed proved too onerous for most students. Teachers wanted students to have more time to work on each piece. By 1989, the NEA reduced the exam requirements, asking students to submit only eight pieces for their exam portfolio, all of which would be assessed. In addition, students did not have to submit separate folders for language and literature. If the eight pieces were appropriately diverse, students would receive two grades for their one portfolio—one for language and one for literature. At least one of the pieces of writing had to be completed under controlled conditions. For the controlled piece, all students in the school wrote on the same topic under test conditions for usually 2 hr. At each school, teachers devised their own controlled writing topics. The NEA (1989) offered guidelines for the controlled writing:

> [It must] represent the candidate's own unaided response to given material and must be done wholly in class. Such assignments should not be preceded by a directed discussion and the material concerned should not be made available, or indicated, to the candidates beforehand. (p. 6)

For the GCSE coursework-only option, the assessment of the writing in the folder was made by the student's teacher, by a committee of teachers in

the school, and was checked and standardized nationally. The national standard setting for portfolio marking in Britain is done somewhat differently by the different examining boards, but the general plans are quite similar. For example, for the NEA, representatives from each school who are teachers and are involved in the national standard setting met twice each year for trial-marking sessions where they received photocopies of scripts or portfolios entered by four students the previous year. The portfolios did not have grades, so the teachers decided 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 Review 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 under this system the students' examination grades for language and for literature were based on an evaluation of the set of pieces in those areas in the folder. As is the case in Vermont, 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.

Under the portfolio-only option, the GCSE was elaborate and standardized, both in the plan for marking the folders and in the plan for collecting the work that went in them. The GCSE shows the crucial role the teacher played in the student's success on a portfolio evaluation, something ultimately the government could not accept. Teachers always play this role, of course, but portfolios place the responsibility unequivocably and directly in the teacher's lap.

CONCLUSIONS

Arts PROPEL, the *PLR,* the Vermont plan, and the GCSE illustrate different ways that portfolio assessment and testing can be used, with the evaluation designs appropriately varied according to the functions they fulfill. In both the U.S. and Britain the large-scale standardized uses of portfolios (the Vermont plan and the GCSE) are running into difficulty. In both cases, the problems center around difficulties implicit in linking

classroom assessment to large-scale testing. In the case of Vermont, the testing program depends on teachers collecting and submitting portfolios; to get teachers to submit portfolios often requires classroom reform that is externally supported. Although Vermont has built that support into its program, the available support is not always sufficient. The NAEP portfolio study also shows the importance of classroom support. Because NAEP provided no advance notice to the teachers, portfolios proved difficult to collect, especially from those teaching in the most challenging situations.

In the case of the GCSE, the difficulties took on a different cast. The problems centered around issues of standardization and fears that the teacher played too much of a role in helping students write the pieces they submitted in their portfolios. Similarly, Gentile (1992) concluded in her report of the 1990 NAEP portfolio study,

> A major concern about using school-based writing samples to assess students' performance for school, district, state, or national assessment purposes is that it is difficult to create the controls necessary to ensure a fair and valid basis of comparison. (p. 68)

NAEP officials also were concerned about interfering with instruction because instruction is part of what NAEP sets out to test.

Among teachers who are developing portfolios in their classrooms, the issues are different still. Calfee and Perfumo (in press) explained:

> (a) teachers who have enlisted in the portfolio movement convey an intense commitment and personal renewal; (b) the technical foundations for portfolio assessment appear infirm and inconsistent at all levels; and (c) portfolio practice at the school and teacher level shies away from standards and grades, toward narrative and descriptive reporting.

For these reasons, the fit between the classroom teacher and the tester, even when portfolios are used by both, is uneven.

Portfolios do, however, fit naturally with good writing instruction and are a potential tool for thoughtful classroom assessment. And portfolios can be used for large-scale testing. The challenge is to make the links. These links between the classroom and the large-scale tester will have to be different from the links usually suggested. For example, Cole (1987) assumed that the assessor is the prime wielder of influence. Given that assumption, she worried that the goals of the teacher and the tester are so disparate that links are impossible; teachers are concerned with instruction, testers with policy and accountability. What she failed to consider is the role that reforms in instruction can play in pushing new kinds of assessment.

Given all the complexities, from the points of view of the tester and the classroom assessor, the hope for linking large-scale testing and classroom assessment will have to lie in a reciprocal relationship. Testers cannot maintain the role of IRS agents, and teachers cannot maintain the role of the audited. Calfee and Perfumo (in press) suggested that if teachers would work together to provide summary judgments of classroom portfolios, then panels could be formed (like the British do for the GCSE) to check individual teachers' judgments and to aggregate the results for testing uses. In this way, classroom portfolios could stimulate and support testing portfolios. In turn, to work together with teachers, testers will have to relax their fears that classroom teachers may in some way contaminate test data collected as a natural part of instruction; after all, it is the teachers' job to influence what students produce. In Vermont, testers, working together with teachers, have benefitted from the classroom portfolio movement and have shown ways that testers can stimulate further classroom reform. But the Vermont experience has also shown the slowness and complexity of the process. Both testers and teachers will have to recognize that the process will be complex and the results of reciprocal relationships will not be immediate. Both teacher and tester will need to enter the dialogue in ways that can be productive for both parties.

ACKNOWLEDGMENTS

This article is adapted from a speech originally presented to the annual meeting of the Chief State School Officers in July 1990. The work reported herein was partially supported under the Educational Research and Development Center Program (R117G10036) as administered by the Office of Educational Research and Improvement, U.S. Department of Education. The findings and opinions expressed herein do not reflect the position or policies of the Office of Educational Research and Improvement or the U.S. Department of Education.

I thank a number of people who were generous with their time in helping me think about and gather information for this article. Bob Calfee provided many ideas, through discussions we had over the years about testing and school reform and more recently as we discussed the potentials of portfolios. Discussions with Anne Haas Dyson provided many ideas as well. From the NAEP, Lynne Jenkins and Ina Mullis answered many questions and helped me gather information about NAEP beyond the published reports. 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 survey of state departments of education. Beth Brenneman provided information about the California Assessment Program. Mary K. Healy made suggestions for the article; Pam Perfumo assisted in gathering materials; Andrew Bouman helped with final editing and formatting.

REFERENCES

Aschbacher, P. (1991). *Alternative assessment: State activity, interest, and concerns* (Tech. Rep. No. 322). Los Angeles: University of California at Los Angeles, Center for Research on Evaluation, Standards, and Student Testing.

Baker, E. (1989). Mandated tests: Educational reform or quality indicator? In B. R. Gifford (Ed.), *Future assessments: Changing views of aptitude, achievement, and instruction* (pp. 3–23). Boston: Kluwer Academic.

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., pp. 241–256). 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* (pp. 44–52). New York: Longman.

Burnham, C. (1986). Portfolio evaluation: Room to breathe and grow. In C. Bridges (Ed.), *Training the teacher of college composition* (pp. 125–138). Urbana, IL: National Council of Teachers of English.

Burstein, L., 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. (1987). The school as a context for the assessment of literacy. *The Reading Teacher, 40,* 738–743.

Calfee, R., & Drum, P. (1979). How the researcher can help the reading teacher with classroom assessment. In L. Resnick & P. Weaver (Eds.), *Theory and practice of early reading* (Vol. 2, pp. 173–206). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Calfee, R., & Perfumo, P. (1992). *A survey of portfolio practices.* Berkeley, CA: University of California, Berkeley, Center for the Study of Writing.

Calfee, R., & Perfumo, P. (in press). Student portfolios and teacher logs: Blueprint for a revolution in assessment. *Journal of Reading.*

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.

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

Cole, N. (1987). A realist's appraisal of the prospects for unifying instruction and assessment. In E. Freeman (Ed.), *Assessment in the service of learning: Proceedings of the 1987 ETS Invitational Conference* (pp. 103–117). Princeton, NJ: Educational Testing Service.

Conlan, G. (1986). "Objective" measures of writing ability. 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.

Cooper, C. R. (1977). Holistic evaluation of writing. In C. Cooper & L. Odell (Eds.),

Evaluating writing: Describing, measuring, judging (pp. 3–31). 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. (1991). Writing. In J. Jensen, J. Flood, D. Lapp, & J. Squire (Eds.), *Handbook of research on teaching the English language arts* (pp. 754–774). New York: Macmillan.

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

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,* 245–255.

Freedman, S. W. (1987). *Response to student writing* (Research Rep. No. 23). Urbana, IL: National Council of Teachers of English.

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

Gentile, C. (1992). *Exploring new methods for collecting school-based writing: NAEP's 1990 portfolio study.* Washington, DC: Office of Educational Research and Improvement.

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, England: Pergamon.

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

Green, B. F. (Ed.). (1981). Issues in testing: Coaching, disclosure, and ethnic bias. *New directions for testing and measurement.* San Francisco: Jossey-Bass.

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

Hewitt, G. (1991). *A little history of Vermont's statewide writing assessment program.* Paper presented at the Albany conference on assessment, SUNY Albany.

Hogan, P. (1974). Foreword. In P. Diederich (Ed.), *Measuring growth in English* (pp. iii–iv). Urbana, IL: National Council of Teachers of English.

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 & L Odell (Eds.), *Evaluating writing: Describing, measuring, judging* (pp. 33–66). Urbana, IL: National Council of Teachers of English.

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.

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.

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

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

Northern Examining Association. (1989). *General Certificate of Secondary Education. English literature syllabus B. Syllabus for the 1989 Examination.* Associated Lancashire Schools Examining Board, Joint Matriculation Board, North Regional Examinations Board, North West Regional Examinations Board, Yorkshire and Humberside Regional Examinations Board: Home Board and Joint Matriculation Board.

O'Connor, M. C. (1989). Aspects of differential performance by minorities on standardized tests: Linguistic and sociocultural factors. In B. Gifford (Ed.), *Test policy and test performance: Education, language, and culture* (pp. 129–181). Boston: Kluwer Academic.

Olmedo, E. L. (1981). Testing linguistic minorities. *American Psychologist, 36,* 1078–1085.

Padilla, A. (1979). Cultural considerations: Hispanic-American. In R. Tyler & S. White (Eds.), *Testing, teaching and learning: Report of a conference on research on testing* (pp. 219–243). Washington, DC: National Institute of Education.

Pelavin Associates. (1991). *Performance assessments in the States* (Report). Washington, DC: Author.

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

Resnick, L. B., & Resnick, D. P. (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* (pp. 63–80). Princeton, NJ: Educational Testing Service.

Rothman, R. (1992, May 20). Vt forced to delay goal of expanding assessment system. *Education Week,* pp. 1, 21.

Samuda, R. J. (1975). *Psychological testing of American minorities: Issues and consequences.* New York: Dodd.

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

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

"This is my best": Vermont's writing assessment program. (Report, 1990–1991). Burlington, VT: The Vermont Department of Education.

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

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

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

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), 35–39.

Wolf, D. P. (1989b, December 13). When the phone rings. *Portfolio: The Newsletter of Arts PROPEL,* p. 1.