RESTRUCTURING LEARNING

1990 SUMMER INSTITUTE PAPERS AND RECOMMENDATIONS

BY THE

COUNCIL OF CHIEF STATE SCHOOL OFFICERS

1993

Evaluating Writing: The Promise of Portfolios as a Link Between Large-Scale Testing and Classroom Assessment

Sarah Warshauer Freedman
University of California, Berkeley

Most English teachers, I suspect, would still agree with the observations that Robert Hogan, then executive director of the National Council of Teachers of English, used to open his preface to Paul Diederich's 1974 book, *Measuring Growth in English*.

Hogan wrote:

> 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.
>
> 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.

The time has come to bridge this rather wide gap between teachers of writing and the testing and measurement community.

To begin to build this new and mutually beneficial linkage, I must first clarify my use of a couple of terms. I am going to use the term "testing" to refer to large-scale standardized evaluation and the term "assessment" to 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." By contrast, Calfee and Drum (1979) note that 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."

It should be acknowledged, however, that this distinction between the two terms is not universal. For instance, by my definition, the business of the National Assessment of Educational Progress (NAEP) is not assessment, but testing. The same is true of the California Assessment Program (CAP).

Accordingly, I plan to focus on two currently distinct kinds of writing evaluation:

- Large-scale testing at the national, state, district, and sometimes school levels; 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.

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. Thus, before presenting some ideas for linking large-scale testing and classroom assess-

Sarah Warshauer Freedman is Professor of Education and Director of the Center for the Study of Writing, School of Education, University of California at Berkeley, California.

ment, I must provide some background about the form of most large-scale writing tests and about the limitations that consequently arise.

Large-Scale Testing

Unlike classroom assessment, large-scale testing generally has not been concerned with charting the development of individual writers. Historically, the large-scale testing of writing was developed to:

(1) Verify that students have mastered writing at some level; the National Assessment of Educational Progress (NAEP) tests fulfill this function;

(2) Evaluate writing programs in the school, district, or in some cases, the classroom; the CAP is one such test;

(3) Place students in programs or classes; and

(4) Select individuals for admission, promotion, or graduation; the Scholastic Aptitude Test (SAT), high school graduation tests, and writing samples gathered by potential employers serve such gatekeeping functions.

Across the years, such large-scale testing programs have struggled with a difficult problem: how to evaluate student writing reliably and cost-effectively.

One highly criticized way to test is through *indirect* measures, which provide proxies for writing abilities. Indirect measures generally take the form of multiple-choice tests and typically ask questions about grammar or logic (for example, the test taker may have to reorder the sentences in a paragraph so that the sequence is logical).

Despite the criticism of indirect measures, they are in widespread use. In 1984, nineteen states measure writing indirectly; only 13 had direct measures, and 18 had no measures at all (Burstein et al., 1985; Baker, 1989).

The appeal of indirect measures is obvious. They're quick to administer and cheap to score. The problems, however, are equally obvious. In the first place, indirect measures are poor predictors of how well the test taker actually writes. In 1986, Gertrude Conlan, a long-time writing assessment specialist at the Educational Testing Service (ETS), commented:

> 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 even indicate whether what the student writes will be interesting to read.

A second negative trait of indirect measures is their negative effect on instruction. If we believe Resnick and Resnick (1990) that you "get what you assess," multiple-choice writing tests discourage instruction that includes writing. If the test does not require writing, why should instruction require it?

From 1890 into the 1960s, the College Entrance Examination Board (CEEB) struggled to find practical ways to move away from indirect, multiple-choice 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; Huddleston, 1954; Meyers, McConville, & Coffman, 1966).

CEEB's struggles were many. First, the student writing had to be evaluated. Apart from the expense of paying humans to score actual writing samples, it proved difficult to get them to agree with one another, even on a single general-impression score. In 1961 at ETS, Diederich, French, and Carlton conducted a study in which "sixty distinguished readers in six occupational fields" read 300 papers written by college freshmen (Diederich, 1974). Of the 300 papers, "101 received every grade from 1 to 9."

In the process, however, valuable insights were gained. On as many papers as possible, the readers wrote brief comments about what they liked and disliked. These comments helped ETS researchers understand why readers disagreed.

By the 1960s, ETS and CEEB had developed ways of training readers to agree independently on holistic (that is, general-impression) scores for student writing, thus solving the reliability problems of direct measurement (Cooper, 1977; Diederich, 1974). For this holistic 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).

In addition to figuring out how to score the writing reliably, the testing agencies 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 tests of writing (Davis, Scriven, & Thomas, 1987; Diederich, 1974; Faigley et al., 1985; Myers, 1980; White, 1985).

When direct-writing tests were relatively novel, the profession breathed a sigh of relief that writing could be evaluated by having students write. Diederich's

opening to his 1974 book typified the optimistic 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.

Unfortunately, however, Diederich's words soon sounded dated. Educators are now raising questions about the validity of the large-scale direct testing of writing.

Many of the tensions center around the unnatural conditions under which tests of writing are taken. Although controlled and written under unaided conditions, as Diederich points out, such writing serves little purpose for students other than to evaluate them. Students must also write on topics that they have not selected and may not be interested in. Another problem is that test takers are not given sufficient time to carry out the elaborate processes that are fundamental to how good writers write and how writing ideally is taught (Brown, 1986; Lucas, 1988a, b; Simmons, 1990; Witte et al., in press). Finally, the evaluation of a student's writing performance is in general based on one or perhaps a few kinds of writing, written in only one kind of context, namely, the testing setting.

The tensions surrounding most large-scale, direct tests of writing are illustrated in the debates currently swirling around the NAEP tests of writing. Every five years, NAEP is supposed to provide "an overall portrait of the writing achievement of American students in grades 4, 8, and 11" (National Assessment, 1990b), as well as track changing "trends in writing achievement" across the years (National Assessment, 1986a).

NAEP gathers informative, persuasive, and imaginative writing samples from students at the three grade levels. For 8th and 12th grade students, 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 6-minute block of background questions common to all students" (National Assessment, 1986a). During a 15-minute block, students write on either one or two topics. For fourth graders, the blocks last only 10 minutes. This means that fourth graders have 5 to 10 minutes to produce up to four pieces of writing during a 30-minute test. Eighth and 12th graders have between 7.5 and 15 minutes to produce up to four pieces during a 45-minute test (National Assessment, 1990a).

Writing researchers and educators who critique the National Assessment argue that we cannot validly make claims about the writing achievement of our nation's schoolchildren, given the NAEP testing conditions, especially the short time students have for writing. To some extent, the NAEP report writers themselves seem to agree (1990b). They 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.

Nevertheless, based on NAEP writing data, *The Nation's Report Card* boldly concludes, "A major conclusion to draw from this assessment is that students at all grade levels are deficient in higher-order thinking skills."

How confident can we be about such a claim?

Indeed, a review of the pedagogical and research literature in writing from the past decade shows an increased focus on a writing process that encourages students to take lots of time with their writing, to choose topics in which they feel some investment, to think deeply before and while they write, and to make use of responses from both peers and teachers as they revise. The NAEP writing test—indeed, most tightly timed test-type writing—is antithetical to current pedagogical trends. What Mellon (1975) said about writing tests 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 [on] 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.

NAEP is attempting to respond to criticisms about the time allotted 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. With increased time, all students scored significantly better on the imaginative section. Fourth and 12th grade students scored significantly better on the persuasive tasks. Only the informative tasks showed no differences. Another intriguing finding is that the extra time proved more helpful to white students than to blacks or Hispanics.

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

> At grade 4, students will be give 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 (National Assessment, 1990a).

Providing 25 or even 50 minutes for writing on a given topic will probably prove insufficient to silence NAEP critics since even that amount of time will not resolve the basic discrepancy between what happens in classrooms and what happens in this testing setting. Therefore, in addition to the double time, NAEP is 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.

Another major point of tension in regard to the NAEP tests centers around scoring. During the mid-1970s, in an effort to obtain more information than a single holistic score and to define clearly the features of writing being judged, NAEP developed an additional scoring system, "the Primary Trait Scoring method" (Lloyd-Jones, 1977).

While the criteria for judging writing holistically emerge from the writing that the students do, the goal of primary trait scoring is to set specific criteria for successful writing on a particular topic before the test is administered. The primary traits are determined and defined by the test maker, who decides what will be essential to writing. Tensions arise, of course, because the test makers cannot always anticipate precisely what test takers will do to produce good writing on a particular topic.

The dilemmas come across clearly through an analysis of Lloyd-Jones's (1977) example of a primary trait scoring rubric. Lloyd Jones explains that one NAEP prompt instructed children: "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." Using primary trait scoring, the writing receives the following scores:

- 0, if the writer gives no response or a fragmented response;
- 1, if the writer does not take a clear position, takes a position but gives no reason, restates the item, gives and then abandons a position, presents a confused or undefined position, or gives a position without reasons;
- 2, if the writer takes a position and gives one unelaborated reason;
- 3, if the writer takes a position and gives one elaborated reason plus one unelaborated reason, or two or three unelaborated reasons; or
- 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 refuses to take 1 position, but instead 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 1 score, but might write a substantially more thoughtful essay than a student who receives a 2, 3, or 4 score after obediently 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 more compelling essay than a student who gives four or more unelaborated reasons and receives a 4.

Indeed, in a study comparing holistic and primary trait scoring, NAEP found that primary trait scoring does not correlate particularly well with the quality judgments; correlations ranged from .38 to .66, depending on the topic (1986a).

Whereas NAEP uses both holistic and primary trait scoring, the latter is the sole means of scoring for the new CAP writing test. The teachers involved in the development of these scoring rubrics report many heated battles, especially given the legislated goal of the test: "to determine the effectiveness of [California public] school districts and schools in assisting pupils

to master the fundamental educational skills toward which instruction is directed" (California Education Code, Section 6061, cited in Loofborough, 1990).

The same issues that have led to the NAEP measurement system also reinforce large-scale, direct-writing assessment at the state level. The same challenges are also present. Several states are meeting the challenges in interesting ways.

One such state is Alaska. Two years ago, 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, 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. Such shortcomings were not lost on Alaska teachers of writing, who are well organized through the Alaska Writing Consortium (an affiliate of the National Writing Project) and who have strong leadership in the state's 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 test that would yield information about students' writing achievement beyond whatever other information that the Iowa test might provide. The state funded an experiment at the 10th grade level, and in 1989, twelve districts participated voluntarily.

The writing was score with an analytic scale, the third commonly used direct scoring method (in addition to primary trait and holistic scoring). The analytic scale offers more information than a single holistic score, but avoids some of the problems associated with primary trait scoring. On the analytic scale, raters give separate scores on ideas, organization, wording, flavor, usage and sentence structure, punctuation and other mechanics, spelling, and handwriting (Diederich, 1974).[1] An analytic scale is used by the International Association for the Evaluation of Education Achievement (IEA) studies of written language (Gorman et al., 1988; Gubb et al., 1987).

For the Alaska test, teachers 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 tests. Thus, students were given a common prompt, but were allowed to work on the writing task during two 50-minute blocks on separate days.

For the Alaska experiment, 60 papers from each of the districts were scored. That sample proved to be enough writing to provide a substantial amount of information beyond what the state board could get from the Iowa test.

In particular, the direct testing showed that knowledge of sentence structure does not guarantee good ideas. The board also learned that direct tests were cost-effective and easy to administer. The number of participating districts has increased to 22 (out of Alaska's 54 districts), and the state's teachers are experimenting with other assessment alternatives as well.

To these alternatives, emerging mostly from the classroom, I will now turn.

New Directions: Writing Portfolios

The portfolio movement provides a potential link between large-scale testing and classroom assessment and teaching. The widespread use of portfolios as evaluation mechanisms could impel important reforms and unite Hogan's accountability agents and the teachers whom they audit.

Portfolios really are not much more than collections of student writing. Mostly classroom-based and designed to provide information about student growth, they have long been a staple of many informal classroom assessments marked by careful teacher observation and careful recordkeeping (for example, anecdotal records and folders of children's work samples).

Through such techniques, changing patterns in behavior over time reveal student progress (British National Writing Project, 1987; Dixon & Stratta, 1986; Genishi & Dyson, 1984; Graves, 1983; Jaggar & Smith-Burke, 1985; Newkirk & Atwell, 1988; ILEA Centre for Language, 1988). Using folders as a basis for discussion, teachers can easily involve students in the evaluation of their own writing—their ways of writing and their products, along with changes in processes and products over time and across kinds of writing activities (Burnham, 1986; Graves, 1983; ILEA Centre for Language, 1988; Simmons, 1990; Wolf, 1988a). Students are thereby helped to formulate concepts about good writing, including the different criteria for good writing according to different situations and audiences (Gere & Stevens, 1985; Knoblauch & Brannon, 1984).

Portfolio uses are not confined to writing classrooms. They are being piloted in a number of other educational evaluation contexts, from mathematics to arts to pilot tests for certifying teachers (through the planned National Board for Professional Teaching Standards).

In a discussion of the uses of portfolios to test teachers, Bird (1988) considers how educators might

borrow the portfolio metaphor from professions like art, design, and photography. Bird contends that such educational uses of portfolios need as clear a definition as they have in other professions, which have conventions that define the nature and contents of a portfolio. In education, there are no such conventions, so according to Bird, the "borrowed idea of 'portfolio' must be reconstructed for its new setting."

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 how writing is collected and used. What writing is to be chosen? Under what conditions is the writing produced? For what purpose? And in what ways is it to be evaluated? Murphy and Smith (1990) suggest another 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 material should be collected?
- What might be done with the portfolios?
- Who hears about the results?
- What provisions can be made for revising the portfolio program?

As the fundamental nature of the questions indicates, portfolio assessment/testing is finding its way into practice well before the concept has been defined. As Wiggins (1990) has pointed out, educators are using portfolios, but their operational definitions range broadly, and their purposes vary widely.

Camp (1990) helped fill in the conceptual gap by listing several essential features for the kinds of writing and thinking activities that will have to be part of portfolios:

- Multiple samples of classroom writing, preferably collected over a sustained period of time;
- Evidence of the processes and strategies that students used to create at least some of those pieces of writing; and
- Evidence of the extent to which students are aware of (1) the processes and strategies they use in writing and (2) their development as writers.

Still, the portfolios that Camp describes involve little more than the collection of actual student work, along with information about students' writing processes and reflections on their work.

To explore the potential of portfolios in large-scale testing, let's look first at how portfolios are being integrated into a school system. Wolf (1988a, b; 1989) writes about Arts PROPEL, a school district portfolio project in art, music, and imaginative writing. Designed as a collaborative involving the Pittsburgh public schools, Harvard University's Project Zero, and ETS, Arts PROPEL aims eventually to provide "alternatives to standardized assessment" (Wolf, 1988b), but first is exploring the power of portfolios to enrich both teaching and learning:

> Central to this [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.

According to Wolf (1989), 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 that 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 lifelong 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 students' development? Conflicts between conventions and inventions?

Wolf stresses 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.

Currently, the Arts PROPEL portfolio data are not used for any assessment purpose beyond classroom

teaching and school-level coordination of information, but as the project resolves the important issues it has raised, the door is opening to the use of portfolios in more generalized forms of evaluation.

Toward Large-Scale Portfolio Use in Schools

How can we begin to link classroom portfolios to assessment and testing goals beyond the classroom? The 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), which was developed in Great Britain.

The PLR is, in effect, a portfolio that keeps systematic records about language growth in all elementary classrooms in the United Kingdom. Written by a committee of teachers and administrators at varied levels, the PLR was piloted and refined in more than 50 schools.

The classroom teacher collects material for the PLR 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; [and] to provide parents with information and assessment of the child's progress" (ILEA Center, 1988).

The British take the position 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 for recording children's language growth—and even standard forms on which to enter some data.

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." Specifically, teachers are to consider:

- The child's pleasure and interest in writing;
- The range and variety of her or 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 or his writing; and
- The understanding the child has of written language conventions and the spelling system.

Teachers are also asked to record observations about children's writing samples at least "once a term or more frequently,"[2] which is thought to be a small extension of previous practice. As the writers of the PLR note, "Many schools already collect examples of children's writing in folders, which become cumulative records"; the sampling that the PLR suggests simply "draws on that practice and allows for the systematic collection and analysis".

In guiding these structured and in-depth looks at particular pieces of student work, the PLR asks for the inclusion of:

- Contextual and background information about the writing;
- The teacher's observations about the child's own response to the writing;
- The teacher's response; and
- Notes on the development of conventions of writing, such as spelling, and what the writing shows about the child's development as a writer.

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

> Ond 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.

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.

The child's response:

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

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.

About the student's developing control of conventions like spelling, 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.

Finally, about the writer's general development, the teacher comments:

> 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.

The PLR is a step forward in that it provides a guide to the teacher for commenting on students' work and for keeping a running record to which others can refer. On the other hand, the PLR, while more specific than any other directives on classroom portfolios, remains relatively vague. For example, the only guidance for the kind of response the teacher might include is:

> 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?

Furthermore, the PLR does not suggest how qualitative comments could be systematically aggregated to provide information about anything other than individual student's development. Certainly, the push to create classroom portfolios has great potential for improving teaching and learning, but such records could also become useful to large-scale testers if we could figure out ways to use the collected data to determine how well students can write and the effectiveness of our curriculum.

Toward Large-Scale Portfolio Use in State Testing Programs

In the United States, we are generally in the experimentation stage with portfolio evaluation systems. We have put them in place at the classroom and school levels in sensible ways, but without worrying too much about their larger uses. However, the hope is (as Wolf wrote) that portfolios will someday replace more traditional forms of large-scale evaluation.

Toward this end, a number of states have begun to support portfolio development work in school settings. Basically, creative teachers and administrators are allowed to experiment with portfolios, tailoring them to local contexts and 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 1st grade classroom in Juneau is using portfolios instead of report cards to determine gains for Chapter 1 programs, and for decisions about promotion to grade 2; and
- Two elementary schoolwide projects are being put in place in Anchorage.[3]

The state of Vermont seems especially advanced in conceptualizing a statewide portfolio program. The Vermont experience shows how assessment and testing goals and classroom reform can be coupled to mutual advantage; however, for now the arrangement is an engagement rather than a marriage—the plan is still only a plan. A draft of the plan, the *Vermont Writing Assessment: THE PORTFOLIO* (Vermont Department of Education, 1989), declares:

> We have devised a plan for a statewide 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.

Vermont plans to assess all students in grades 4 and 11 according to the state's three-part plan. First, students will write one piece to an assigned and timed prompt. The result will be scored holistically. Second, with the help of their classroom teacher, students will select and submit a "best piece" from their classroom writing portfolio. The same teachers who evaluate the prompted sample will score the portfolio selection. Finally, state evaluation teams will visit all schools "to review a sample of 4th and 11th grade portfolios." 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." The scores from the prompted sample and the best piece are supposed to indicate each student's writing abilities, while the sample portfolios will provide a picture of the school's writing program.

For the classroom portfolios, the draft of the Vermont plan advises students to keep all drafts of any piece the student might want included. The plan also advises schools to buy or clear storage cabinets so that students can keep a full current-year folder. Part of the current-year folder's contents will later be transferred to a permanent folder, which will contain selections of the students' work from grades kindergarten through 12.

The current-year folder is to contain a cover sheet much like the one in the PLR. It will have space for teacher comments, instructions and goals for students; the state evaluation team's official comments; and a combination grid and checklist for documenting the process of producing the portfolio work.

The state team is likely to recommend a minimum set of pieces of varied types for inclusion in the portfolio, either (1) something expressive, imaginative, informative, persuasive, and formulaic (something written to fulfill social obligations or a letter explaining the choices of the work in the portfolio), or (2) 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 evaluating the portfolio stipulates, "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." The questions include both a scale with a numerical score and a place for qualitative comments. For example, Figure 1 shows the first of the 14 scaled questions.

Other questions for the teacher-evaluators ask about audience awareness, logical sequence, syntax, and spelling; the process the student used to produce the pieces and the folder; and the coherence of the folder as a whole. The qualitative comment section is similar to but simpler than that in the PLR—it provides only spaces for general observations and recommendations.

The comprehensive Vermont plan provides for teacher inservice in the collection and evaluation of student portfolios as well as for a statewide evaluation of 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.

Figure 1
Sample Portfolio Evaluation Question

[ ]* 1. DOES WRITING REFLECT A SENSE OF 2 3 4 5 6 7 8**

AUTHENTIC VOICE?

[ ] Somewhat [ ] Consistently [ ] Extensively***

Notes:

* = Check box (informal)

** = Holistic score (formal)

*** = Graduated terms (informal)

Although Vermont's plan is still in the developmental stages, the state seems to be leading the way in connecting teacher inservice 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.

Toward Large-Scale Portfolio Use for National Examinations

As a final example of ways to link classroom assessment and widescale testing, I would like to describe the national examination that determines whether or not British students at age 16 or older 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).[4]

The GCSE serves a major gatekeeping function in Great Britain. First, British students must receive high scores on the GCSE to go on to a two-year course, the General Certificate of Education at Advanced Level, popularly known as *A levels*. The A-level courses qualify students for entry to universities and other forms of higher education, and some employers demand A levels. Over 60 percent of U.K. students do not take A levels, but leave school after taking the GCSE examination.

For the GCSE in language and literature, schools choose between either (1) a timed examination at the end of the two years plus a folder of coursework (portfolios) or (2) simply the folder of coursework. In the case of English language and literature examinations, the coursework included in the portfolio is student writing. For the GCSE examination, schools have a choice from an examination syllabus, which offers different options in the format and organization of the examination.

The specifications for the GCSE differ slightly according to the five examining boards in England and Wales. Whatever the variations, however, the important point is that the GCSE is now a national, large-scale examination based in large part or entirely on portfolios of students' coursework.

For the coursework-only option, students must complete 20 pieces of writing, 10 for the English language examination and 10 for the literature examination (the two examinations are assessed separately). The writing samples in the folder must be in a variety of genres, for a variety of purposes, and for different audiences. Items may include, for example, reports, description, argument and persuasive writing, narrative fiction, poems, and responses to text.

Such samples are assembled over a two-year period (usually with the same teacher for both years of the examination course. Of the 10 pieces for each examination, the student and teacher choose the 5 best pieces that cover the assessment objectives for each examination. These are the pieces that are finally evaluated.

For the coursework-only option, the student's teacher and a committee of other teachers in the school assess the writing in the coursework folder. The work is also 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. Representatives from each school (all are teachers who are involved in the national standard setting) meet twice a year for trial marking sessions. They receive ungraded photocopies of portfolios entered by four students the previous year. After the teachers decide the grades they would give if the candidates were their students, the teachers present their grades at a school meeting. There, the portfolios are discussed and school grades agreed on.

Representatives from each school attend a consortium trial marking meeting at which portfolios and grades are discussed once more. A member of the National Review Board of the Northern Examining Association (NEA) attends this meeting and explains the grades that the board has given.

After this training period, a committee of teachers in the school agrees on grades for the coursework folders from their 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 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 students examination grades for language and for literature are based on an evaluation for the set of pieces in the folder. The portfolio evaluation consists of a grade given for a group of pieces; it is not derived from an average of grades on individual pieces.

All assessors, including the National Review Panel, are practicing teachers, which is a very important attribute of the system. If we are going to make decisions about students based on their coursework folders, teachers have to play a crucial role in the students'

success. Teachers have always played an important role, of course, but portfolios place a greater degree of assessment responsibility unequivocally 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 that the assessments are intended to fulfill. Although the models of well-conceived, large-scale portfolio programs are few, they are certainly beginning to emerge, and they are remarkable for their thoughtful approach to students and to the evaluation of their written work.

Conclusion

In the assessment of writing, the concept of the portfolio seems particularly appealing because writers, like artists, collect representative work samples that provide a sense of the range and quality of what individual students can do (Anson, Bridwell-Bowles, & Brown, 1988; Burnham, 1986; Camp, 1985a, b; Camp, 1990; Elbow & Belanoff, 1986a, b; Fowles & Gentile, 1989; Lucas, 1988a, b; Murphy & Smith, 1990; Simmons, 1990; Stiggins, 1988; Wolf, 1988a, b).

Portfolios can be collected as part of an ongoing instructional program, which can get around the problem of one-shot evaluation procedures (Anson et al., 1988; Belanoff, 1985; Burnham, 1986; 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, 1988a).

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, 1981; Cooper & Murphy, in progress; Cooper & Odell, 1977; Diederich, 1974; Loofborough, 1990; Mellon, 1975; Myers, 1980; Resnick & Resnick, 1977, 1990). There is, therefore, an important role for teacher-driven and classroom-based assessment in our plans for educational reform.

I want to end, however, with a word of warning. Currently in the United States, the National Assessment of Educational Progress 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 that I have described, the writing that is submitted is rather dismal.

Just collecting and evaluating portfolios will solve neither our assessment problems nor our need to create a professional climate in our schools. As the careful work of the Pittsburgh Arts PROPEL project shows, however, coupling assessment and instruction in increasingly sophisticated ways presents us with a highly promising opportunity to make a real difference in education in this country.

Endnotes

[1]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.

[2]In the United Kingdom, the school is divided into three terms: Fall, Winter, and Summer.

[3]Other states implementing or experimenting with portfolio assessments 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. Pamela Aschbacher of UCLA's Center for the Study of Evaluation compiled this information through telephone interviews with officials in each state's department of education.

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

REFERENCES

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

Belanoff, P. (1985). Models of portfolio assessment. In K. L. Greenberg & V. B. Slaughter (Eds.), *Notes from the National Testing Network in Writing.* 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, 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, 8,* 438–443.

Calfee, R., & Drum, P. (1979). *Teaching reading in compensatory classes.* Newark, DE: International Reading Association. (Eric Document No. ED172179)

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.

Calkins, A. (1990). Personal correspondence.

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

Charney, D. (1984). The validity of using holistic scoring to evaluate writing: A critical overview. *Research in the Teaching of English, 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.* 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. (1981). Competency testing: Issues and overview. In C. R. Cooper (Ed.), *The nature and measurement of competency 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 classroom." Unpublished manuscript.

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

Diederich, P., 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.

Elbow, P. (1986). Portfolio assessment as an alternative in proficiency testing. In K. L. Greenberg & V. B. Slaughter (Eds.), *Notes from the National Testing Network in Writing.* New York: The City University of New York Instructional Resource Center.

Elbow, P., & Belanoff, P. (1986a). Portfolios as a substitute for proficiency examinations. *College Composition and Communication, 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. (1981). Influences on evaluators of expository essays: Beyond the text. *Research in the Teaching of English, 3,* 245–255.

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.* Norwood, NJ: Ablex.

Godshalk, T., Purves, A., & Degenhart, R. (1988). *The IEA study of written composition I: The International Writing Tasks and Scoring Scales.* Oxford: Pergamon Press.

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

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

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

ILEA Centre for Language in Primary Education. (1988). *The Primary Language Record: Handbook for teachers.* London: ILEA Centre for Language in Primary Education.

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 Monclair, NJ: Boynton/Cook.

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

Loofborough, 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 Ph.D. dissertation, University of California-Berkeley.

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

Lucas, C. K. (1988b). Toward ecological evaluation. *The Quarterly of the National Writing Project and the Center for the Study of Writing, 1,* 1–3 and 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,* 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,* 1–3 and 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.* Princeton, NJ: Educational Testing Service.

National Assessment of Educational Progress. (1990a). *Learning to write in our nation's schools.* 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.), *Process, development, and communication. Vol. 2. Writing: The nature, development, and teaching of written communication.* Hillsdale, NJ: Erlbaum.

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

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.* 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, 3,* 262–268.

Stiggins, R. J. (1988, January). Revitalizing classroom assessment: The highest educational priority. *Phi Delta Kappan, 69*(5), 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 Department of Education. (1989). *Vermont Writing Assessment: THE PORTFOLIO.* Montpelier: Vermont Department of Education.

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

Wiggins, G. (1990, January 24). "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. (1988a). Opening up assessment. *Educational Leadership, 45*(4), 24–29.

Wolf, D. P. (1988b). Portfolio assessment: Sampling student work. *Educational Leadership, 46*(7), 4–10.

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