Portfolios, Accountability, and an Interpretive Approach to Validity

Pamela A. Moss, Jamie Sue Beck, Catherine Ebbs, Barbara Matson, James Muchmore, Dorothy Steele, Caroline Taylor, *University of Michigan*, and Roberta Herter, *Henry Ford High School*

How can the results of classroom-based portfolio assessment be communicated outside the classroom? How might a portfolio-based assessment system be designed and implemented? How can we evaluate the merits of portfolio-based assessments?

In our current educational milieu, information about student learning is often required to serve accountability purposes—to provide evidence to parents, administrators, policymakers, and other concerned citizens that students are receiving high quality education. Frequently, the information provided is based upon externally imposed assessments over which teachers and students have little or no control. Assessment information of this type may have little relevance to the classroom context, and yet pressure to do well—perceived or real—gives these assessments substantial influence over decisions about instruction and learning.

In this article, we focus on how portfolio-based conclusions about student learning might be used to communicate in credible ways with various audiences outside the classroom, thus serving some of the same accountability purposes that externally imposed standardized tests typically serve. First, we address the issue of why it is important that information from classroom-generated portfolio assessments be used to communicate with various stakeholders outside the classroom—why such contextualized information about student learning should be used to balance information pro-

vided by externally imposed performance assessments. Second, we suggest alternative procedures and criteria for investigating validity in these less standardized domains—procedures and criteria that draw on interpretive rather than empirical analytic or positivist research traditions. Finally, using portfolios from an eighth-grade Language Arts classroom, we offer an example of how portfolio-based conclusions about student achievement and growth might be used to communicate with a broader audience in well warranted and credible ways.

Why Use Classroom-Generated Assessment for Accountability Purposes?

The Impact on Teaching and Learning

Over the past decade, it has been repeatedly demonstrated that assessment influences what students learn and what teachers teach. In a recent review of literature on the impact of classroom evaluation on students, Terence Crooks (1988) concluded:

> Classroom evaluation ... guides [students'] judgment of what is important to learn, affects their motivation and self-perceptions of competence, structures their approaches to and timing of personal

study, consolidates learning, and affects the development of enduring learning strategies and skills. It appears to be one of the most potent forces influencing education. (p. 467)

Similar conclusions have been drawn about the impact of district- and state-mandated assessment on teachers' judgments concerning what to teach and how to teach it. Because of the widespread public attention that standardized test scores receive and the tangible rewards and sanctions frequently associated with student performance, district- and state-mandated tests have become instructional magnets, providing potent statements to teachers and administrators about what is important for students to learn (Madaus, 1988; National Commission on Testing and Public Policy, 1990; Resnick & Resnick, in press; Smith, 1991). As Resnick and Resnick (in press) conclude in their review of literature on the impact of high-stakes testing on instruction, over time high-stakes tests tend to drive out other educational concerns, narrowing the curriculum to the objectives reflected on

Pamela A. Moss is an assistant professor in the School of Education at the University of Michigan, 610 E. University, 4220 SEB, Ann Arbor, MI 48109-1259. She specializes in educational measurement and evaluation.

Jamie Sue Beck, Catherine Ebbs, Barbara Matson, James Muchmore, Dorothy Steele, and Caroline Taylor are graduate students in the School of Education at the University of Michigan.

Roberta Herter is an English teacher at Henry Ford High School in Detroit, MI.

the tests. Consequently, it becomes crucial to assess the full range of knowledge, skills, and interests we want to nurture in our students and their teachers.

In recent years, educators have argued that performance assessment—assessment involving extended discourse, work exhibits, or other performances—is essential in nurturing critical and creative thought (e.g., Johnston, 1989; Newmann, 1990; Wolf, Bixby, Glenn, & Gardner, 1991). A number of districts, states, and national committees have begun to explore the feasibility of using large-scale, standardized performance assessments to complement or replace standardized multiple-choice assessments as a means of educational accountability and reform. While we fully endorse the importance of performance assessment in nurturing critical and creative thought, we are concerned that exclusive use of standardized performance assessment limits learning opportunities in significant ways.

Implicit in typical models of assessment serving accountability purposes is the need for centralization of authority within a given context to decide specifically what is measured and how it is measured; tasks, scoring procedures, and administration conditions are standardized in order to enhance comparability of scores from task to task, scorer to scorer, and subject to subject. This model, which separates assessments "that count" from ongoing day-to-day activities, has been criticized for disenfranchising teachers and students from setting their own intellectual problems and from debating the criteria and standards that will be applied to their work. The potential negative impact of such a model has been noted in the context of both student development (e.g., Camp, 1992; Johnston, 1987; Willinsky, 1990) and teacher development (e.g., Darling-Hammond, 1989; Johnston, Weiss, & Afflerbach, 1990; Smith, 1991).

The assessment model we propose is one in which teachers and students are encouraged to collaborate with one another to make intellectual and creative choices consistent with their own goals and interests, to engage in ongoing reflection about

their work, and to participate in the development of strategies by which achievement and development are shared with various stakeholders outside the classroom. This approach is largely consistent with what Darling-Hammond (1989) calls a professional model of accountability, which seeks evidence that teachers are making knowledge-based decisions that respond to individual student needs. (Externally imposed assessments, by contrast, are more consistent with what Darling-Hammond calls a bureaucratic model of accountability, which seeks evidence that teachers are adhering to standardized policies and procedures for groups of students.) This call for professionalization of teaching, emphasizing practice that is student centered and knowledge based, is consistent with a number of educational reform efforts, including those of the Holmes Group (1986) and the National Board for Professional Teaching Standards (1990).

The Quality of the Information Provided
With few exceptions, performance assessments that serve policy purposes at the district, state, and national level involve one or two standardized prompts to which students must write responses in a few minutes to a few hours at best. Although these writing performance assessments do, to some degree, assess and encourage critical or creative thought, they rarely provide an opportunity for students to engage in much of the process of writing, especially the rethinking and revising typical of the way that experienced writers work. Nor do they encourage students to discover their own purposes for writing beyond showing their competence in school (Johnston, 1987; Newmann, 1990). The objectives and prompts are necessarily homogenized to accommodate a variety of different curricula and, as such, may not adequately represent the instructional content of any particular curriculum or the interests and goals of individual students. Scoring is centrally done, based upon standardized criteria by readers who have little or no knowledge of the students, the goals they or their teachers have set, the learning oppor-

tunities that they were provided, or the criteria that they and their teachers value. Typically, once each individual piece of writing is scored, human judgment about the writing itself stops, and conclusions about achievement and growth in writing are based on statistical manipulations—accumulation or comparison of scores from parallel or equated items. Neither achievement nor growth can be noted in areas not addressed by the standardized scoring criteria.

Our experience is that growth is often manifest in qualitative changes in the writing—changes in the complexity of the problems that students undertake, which may involve losing control over other features of the writing like organization or mechanics. Take Gretchen, a student from the classroom described in our example below, who included two pieces of expository response to literature in her portfolio. In one sense, the second piece is not as strong as the first—it is not as well organized or coherent—but it is a richer interpretation. Unlike the first piece, which simply compares two groups of characters from *Lord of the Flies* on a variety of features, the second piece, on *Animal Farm*, has a thematic framework about the role of scapegoats that is played out with evidence from Gretchen's own personal experience, from the novel, and from a definition of the term acquired from another resource. A comparison of Gretchen's revisions in the two pieces shows a newly developed awareness of the need for elaboration and for evidence on particular points.

Comparing holistic scores on expository response to literature would not pick up these subtle changes in Gretchen's writing. Even comparing a series of analytic scores that might, for example, show increases in quality of ideas and voice with decreases in focus, organization, and mechanics would not adequately represent Gretchen's growth, unless *patterns* of scores for each piece of writing were compared over time. To average the scores from the two pieces so as to talk about Gretchen's achievement in expository writing or to subtract or otherwise manipulate the scores to talk about growth, would miss the point. Interpretation

based on complex comparisons of the two writing samples is crucial in discovering Gretchen's achievement and growth as a writer.

Consequently, we suggest an assessment model in which teachers and students are encouraged to make intellectual and creative choices that reflect their own goals and interests and in which teachers' interpretations of their students' work—individually and as a class—play the central role. Writing portfolios are particularly valuable in promoting and assessing student learning because they contain rich and varied examples of student work, collected over extended periods of time. They may also include multiple drafts of work, reflecting the ways in which students use the processes of writing to refine ideas; students' self-assessments, reflecting their conceptions of themselves as writers; and the teacher's and others' responses to the student work, reflecting classroom-generated criteria and standards of writing. Narrative interpretations by the teacher can place the student work within the context of learning opportunities available in the classroom and student experience and interests. These narrative interpretations, whether oral or written, supported by actual samples of student work, are more likely to facilitate genuine understanding and dialogue among those using and those providing assessment information than are frequency distributions of standardized test scores. With standardized test scores, consumers of the information must rely on expert assertions about the meaningfulness of the results because they do not have access to the evidence on which the conclusions were based. With well-documented narratives, all stakeholders have access to at least samples of the evidence supporting the conclusions, and genuine dialogue about the meaningfulness of the conclusions can take place.

An Interpretive Strategy for Drawing and Warranting Conclusions

How do we validate teachers' interpretations of their students' achievement and growth in writing—interpretations based on work samples that may vary substantially from student to student and from class to class? Traditional criteria for assessments have little to offer us, at least in the form by which they are typically operationalized, because of their dependence on standardization to enhance reliability and to permit comparison and aggregation of data.

Certainly the issues underlying our concerns about reliability are relevant in the portfolio context as well—questions about what we can validly and fairly conclude from particular samples of writing evaluated by particular readers. Without standardization, however, traditional criteria of reliability and comparability are likely to be unacceptably low. And yet, as suggested above, a case can be made for the enhanced quality of information that includes an integrative interpretation of the achievement and growth reflected in student work, based upon an intimate knowledge of the learning context. As one possibility for warranting such interpretations, we suggest that teachers and others who might use portfolios treat the assessment process as an instance of interpretive or qualitative research, drawing upon the epistemological strategies typically used by interpretive researchers to develop, warrant, and present their conclusions. In describing these strategies, we draw heavily upon the language of Erickson (1986), although the general approaches we describe are consistent with the methodological writing of other interpretive researchers in education (e.g., Goetz & LeCompte, 1984; Lincoln & Guba, 1985).

The analytic process that Erickson and other interpretive methodologists describe is largely inductive and iterative. The process typically involves the development of a coding scheme—a set of categories that link various items of data—which is used in developing and articulating assertions or conclusions. Preliminary assertions are repeatedly tested against the entire body of data, as both confirming and disconfirming evidence is sought. Assertions are revised until they can account for all the data presented. The manner in which the report is written enhances the credibility of the conclusions by allowing the reader to serve as co-analyst. Erickson (1986) suggests that the research report should contain *particular descriptions*—quotations or narrative vignettes—that illustrate the meaning of different categories and conclusions; *general descriptions* that provide evidence of the relative frequency of a given phenomenon and display the breadth of evidence that warrants the conclusion; and *interpretation* that frames the particular and general descriptions, tying them to the conclusions. Thus, the main aim in writing the research report is "to persuade the audience that an adequate evidentiary warrant exists for the assertions made" (p. 148). Among the additional strategies typically suggested to enhance the validity or credibility of conclusions are long-term involvement by the researcher in the specific research context; persistent observation; triangulation across data sources, analytic methods, and researchers; trying out working assertions with a disinterested peer; and leaving an audit trail so that evidence supporting the conclusions can be reviewed by others (Lincoln & Guba, 1985). (All these strategies might be put to good use in the context of portfolio assessment within a professional model of accountability.)

To illustrate how this type of interpretive approach might work in practice, we offer what is as yet a primitive example of how portfolio-based interpretations of student work might be developed, reported, and evaluated. In offering this example, we want to acknowledge at the outset that we see our role in this endeavor as somewhat artificial. Our goal was to suggest an approach that teachers might use to communicate with various outside audiences in well warranted and credible ways. The willingness of one teacher and her students to share their work gave us an opportunity to develop a possible strategy. Ultimately, however, we view the role of outsiders as co- or secondary analysts, who might sample portfolios and narratives to trace the path of a teacher's argument and raise alternative interpretations for consideration where appropriate, but as secondary to the interpretive process. The procedures we suggest here address accountability concerns that focus on the progress of individual students and the professional judgment of teach-

ers. We also offer brief suggestions for how program, school, district, or state level information might be constructed when accountability concerns require system level information.

Data Source

Our example draws on portfolios from the eighth-grade Language Arts classroom of Kathryn Howard, a teacher at Reizenstein Middle School in Pittsburgh, Pennsylvania, and a member of the Arts PROPEL team. Arts PROPEL, a Rockefeller funded collaborative, involves teachers and researchers from the Educational Testing Service, Harvard Project Zero, and the Pittsburgh Public Schools in developing portfolio assessments closely related to instruction in music, visual arts, and imaginative writing in grades six through twelve (Camp, 1992; Wolf, 1989). In English/Language Arts, the PROPEL project has expanded to encompass expository as well as imaginative writing (Howard, 1990).

Consistent with her goals of encouraging students to make choices that reflect their own goals and interests and to observe the effects of those choices over extended periods of time, portfolios in Howard's class were "owned" by students. Periodically, throughout the year, students selected pieces from their comprehensive writing folders to include in their portfolios, thus providing a portrait of themselves as developing writers. They were encouraged to select an "important" piece, a "satisfying" piece, an "unsatisfying" piece, and a "free pick"; to include multiple drafts of each piece; and to reflect, in writing, on their reasons for selection, the strengths of the writing, and their goals for subsequent work. Howard's reflections on these pieces, also included in the portfolios, typically focused on one or two things done well and one or two things to work on in the future. Students were also encouraged to include three more general self-reflections about their development as writers—at the beginning of the year, in the middle of the year (in response to parents' comments about their writing folders), and at the end of the year. The writing in Howard's class spanned genres of poetry, narrative, drama,

expository response to literature, and personal writing. Because students were free to select from any of these genres in creating their portfolios, the contents of the portfolios varied substantially from student to student.

Creating Narrative Profiles for Individual Students

We began by creating an interpretive framework—a coding scheme, if you will—that could be used to analyze the data contained in the portfolios. We worked inductively, from the portfolios, writing folders, and the other material described above, and deductively, from our own knowledge and values and from our readings about the genres and purposes of the writing we encountered, in order to develop a list of features to be used in analyzing the contents of the portfolios.

We decided to treat the "sequence" of writing as our unit of analysis. A sequence, as we defined it, consisted of a finished product or final draft, plus all related preliminary drafts or plans, the student's self-reflections about the writing, the teacher's reflections about the writing, and the reflections of others who might have responded to the student's work, typically parents or peers. Each portfolio contained three to five sequences of writing in addition to the three general self-assessments.

Our interpretive framework, which could be applied to each sequence of writing in a portfolio, is depicted in Figure 1. To illustrate its use, we have duplicated our codes and comments for a sample sequence of writing by Barry, one of Howard's students. Figure 2 contains the actual sequence of writing: Barry's final draft of an expository response to *Lord of the Flies,* a preliminary draft, his self-reflection about the piece, and Howard's comments to Barry. The first column in Figure 1 presents the list of features that we developed to describe each sequence of writing, divided into five clusters. *Vision* refers to the reader's reconstruction of the writer's semantic intent—the message apparent in the writing; *development* refers to the means by which that message is conveyed and can be seen in the interrelationship among different as-

pects of the writing; *craft: language and form* refers to various standard conventions of language and form; *craft: literary style* refers to the nonordinary use of language that goes beyond the conventional meaning conveyed in the words on the page; *reader's response* refers to our own cognitive and emotional response to the writing; and *sense of writer* refers to what the writer or others explicitly tell us about the writer.

For each sequence of writing, we recorded the genre, the date it was completed, the writer's reason for including it in the portfolio, and the numbers of drafts, self-reflections, and comments by the teacher and others. Observations on the sequence of writing are recorded in the subsequent columns of Figure 1, which refer to the components of each sequence. The column labeled *final* refers to the finished product— the piece that we assumed the writer intended to share. In this column, we recorded the extent to which each of the features on our list was evident in the writing, ranging from "no evidence of feature" through "integrated, sophisticated evidence of feature." The column labeled *revise* refers to changes evident in editing marks or in comparing the final draft to earlier drafts. Here, we indicated which features of the writing the revision affected, the extent of the revision, and whether or not it increased or decreased evidence of use of that feature. The columns labeled *self, teacher,* and *other* refer to reflections about the writing by the student writer, the teacher, or other readers (e.g., parents, peers) respectively. Here, we indicated the features of the writing or the writer which the reflection addressed and whether the features were positively, negatively, or neutrally evaluated. The *comments* column leaves room to note features of the student's work or reflections by others that the framework could not accommodate.

It is important to emphasize that our intent was not to "score" the student's writing. A strong piece of writing, within a particular genre, might well have low evidence of use of a number of these features. For example, even in a strong expository response to literature, one is un-

	FINAL	REVISE	SELF	TEACHER	OTHER	COMMENTS
date	1/18/90					
length/number	1 page	4 drafts	1	1	0	
VISION						
something to say	3			/		T: explain, why would you rather be with Ralph, not Jack?
voice	2					
seeing beyond ordinary	2					
conceptual framework	2					
autobiography	1					
intertextuality						
DEVELOPMENT						
elaboration	3					
specificity	3					
explanations	3			-		
focus	2					
structuring	3	+		+		
comprehensiveness	2			-		T: you confuse Ralph & Jack in 2nd paragraph
movement	3	+		+		
multiple perspectives	1					
CRAFT: LANGUAGE/FORM						
word choice	3	+	-			S: I'd like to bring in more fancy words
sentence structure	3	+				
clarity/cohesiveness	3	+				
form approp. to genre	3	+				
standard conventions	3	+				
CRAFT: LITERARY STYLE						
sound						
figurative language						
concrete imagery	1					
shape						
show-not-tell	1					
tone						
READER'S RESPONSE						
compelling/interesting	2					
believable	3					
accessible	3					
shared context	3					
SENSE OF WRITER						
engagement/interest			/			
comfort						
gen'l self-evaluation						
purpose awareness						
genre awareness			+			S: It's not like creative writing It's more concrete
audience awareness						
use of comments						
revising						
development						

FINAL
1 some evidence of feature
2 extended evidence of feature
3 appropriately sustained evidence of feature
4 integrated/sophisticated evidence of feature

REVISE
+ revision increases evidence of feature
/ revision doesn't change evidence of feature
- revision reduces evidence of feature

single sign = some revision
double sign = extended revision

SELF, TEACHER, OTHER
+ positive mention
/ neutral mention
- change suggested

single sign = some mention
double sign = extended mention

FIGURE 1. _Completed framework for Barry's sequence on_ Lord of the Flies

likely to see evidence of play with the "sound" of language; similarly, a compelling poem might show little evidence of elaboration. Rather, our intention was that the frameworks, one for each sequence in the portfo-

There are a number of striking differences between Jack and Ralph. First, Jack always took the violent and outrageous approach to everything. However, Ralph, instead of taking a brutal approach, tried to think of peaceful solutions to problems. For example, when Ralph wanted to build huts for the boys to live in, Jack wanted to go out and hunt. Another example occurs near the end of the book, when Ralph was trying to get everybody rescued, while Jack was running around in circles singing brutal chants like "Kill the pig. Cut her throat. Bash her head in."

These incidents show another difference between Jack and Ralph. Ralph was really more mature than Jack. Jack would rather hunt than sit down and plan things out, because hunting was more fun than thinking and planning. But Jack understood that if the boys were going to survive, they had to do much more than hunt and have fun.

Both Ralph's and Jack's tribes were becoming more savage and brutal as the book went on. Even though Ralph was much more peaceful than Jack, he was having trouble surviving without violence. The cause of Ralph's violence was Jack's greater violence. Ralph had to compete with Jack. Ralph's violence was in self-defense.

Thus, in summary, there were important differences between Jack and Ralph. If I had to be marooned on a desert island, I would rather be with Ralph than with Jack.

• Preliminary Draft for Barry's Sequence

There are a number of striking differences between Jack and Ralph. First, Jack always took the violent and outrageous apprroach to everything. At this point Ralph, insted instead of taking a wonton approach, Rafe Ralph tried to think it out first and solve problems.

¶ For example, when Ralph wanted to build huts, Jack wanted to go out and hunt. Another examples is towards the end of the story BOOK Ralph was trying to evarybody get rescued, while Jack was working out with the tribe by RUNNING IN CIRCLES singing BRUTAL song CHANTS, "Kill the pig." "Cut her throat." "Bash her in." Ralph was also more mucure than Jack. While Jack was hunting it was funner FOR HIM TO HUNT THAN TO AACTUALLY SIT DOWN AND PLAN THE WHOLE THING OUT.

¶ Both Ralph's group and Jack's tribe were both becoming more savagelike, or brutal, but towards the end Ralph was having fu trouble surviving without violence. ¶ Ralph would have to chang to violence. The cause of this would be because jack was very violent and Ralph had to compete in self defence Thus, in summary, there are many diffrences between Ralph and Jack, so, if I was marooned on a desert Island, I rather be with Ralph, then, Jack.

• Self-Reflection for Barry's Sequence

Please describe the writing assignment (2-3 sentences explaining what you were asked to do). You had to compare or contrast Ralph and Jacks tribe or just Jack and Ralph.

What do you like best about this piece of your writing? I like that it is not creative writing. It is more concrete.

What would you like to continue to work on in future writing? I would like bring in more a fancy words in my pieces.

• Teacher's Reflection for Barry's Sequence

One thing that is done well:
Your transition from first to second paragraph is simply excellent. It enhances the paper's flow.

One thing that needs to be improved:
In the second paragraph you confuse Ralph and Jack.
Also, why would you rather be with Ralph? Explain.

FIGURE 2. *Barry's sequence of writing on* Lord of the Flies

lios, would help the portfolio readers *describe* students' work—to attend to the entire contents of the portfolios, to abstract relevant information, and to record it for later use. In all cases, the narratives were written by the person who read the portfolios and filled out the frameworks. The frameworks themselves were only intermediate steps, undertaken to inform the writing of the narrative profiles.

After completing an interpretive framework for each sequence in a student's portfolio, we prepared a

Date	Genre	Topic	Reason	Length	Drafts
9/??	Self-Reflection	Thinking About Your Writing	Requested	1 page	1 draft
10/17	Narrative/Dramatic	Personal Monologue	Important	1 page	2 drafts
1/16	Response to Literature	On The Lord of the Flies	Unsatisfying	1 page	4 drafts
2/??	Self-Reflection	Response to Parent Comments	Requested	1 page	1 drafts
2/28	Narrative/Dramatic	"The Tell-Tale Heart	Free Pick	3 pages	2 drafts
5/22	Response to Literature	On Animal Farm	Satisfying	5 pages	2 drafts
6/??	Self-Reflection	Final Reflection	Requested	2 pages	1 draft

As a writer, Barry shows substantial growth from the beginning of the year in his first personal monologue to his last piece, a response to Animal Farm . Initially, Barry seems to have little control over the flow and transition of his ideas. His points are not tied together, he jumps around in his thinking, and he lacks specificity in his ideas. By January, when Barry writes his response to The Lord of the Flies, he begins a coherent argument about the differences between Ralph's group and Jack's tribe, although he ends with the unsupported assertions that he would have preferred to be "marooned on a desert island" with Ralph. Barry includes three reasons for his comparison, hinges his reasons with transition words, but more impressively, connects his introductory paragraph with a transition sentence to the body of his essay. In the revisions of this essay, Barry makes primarily word and sentence level changes, adds paragraph formatting, and generally improves the local coherence of the piece.

By the end of February when he writes his narrative response to Poe's "The Tell-tale Heart," Barry displays a concern for making his writing interesting. "I like the idea that there are so many twists in the story that I really think makes it interesting". He makes surface level spelling changes, deletes a sentence, and replaces details, although not always successfully (e.g., "fine satin sheets and brass bed," is replaced with the summary description "extravagant furniture"). Overall, it is an effective piece of writing showing Barry's understanding of narrative form and his ability to manipulate twists of plot in order to create an engaging story.

Barry's last selection in his portfolio is an exceptional five page, typed essay on Orwell's Animal Farm. The writing is highly organized around the theme of scapegoating. Using supporting details from the novel and contemporary examples from politics and sports, Barry creates a compelling and believable argument. The effective intertextuality and the multiple perspectives Barry brings to this essay result largely from an exceptional revision process. Not only does he attempt to correct his standard conventions and improve his word choices; he also revises successfully to the point of moving around whole clumps of text and adding sections that significantly reshape the piece. This pattern of revision shows the control Barry has gained over his writing.

In Barry's final reflection he describes his development, showing an awareness of such issues as organizing and connecting ideas, choosing appropriate words and details, and making his writing accessible to his readers. "I had many gaps in my writing. One problem was that I would skip from one idea to the next and it would not be clear what was going on in the piece....Now, I have put in more details so you don't have to think as much as you would. I also perfect my transitions and my paragraph form.... My reading...has improved my vocabulary and it helped me organize my writing so it sounds its best and makes the most sense possible.... There are many mistakes I have made throughout the year, but I have at least learned from all of them." I agree with him.

FIGURE 3. Narrative profile for Barry's portfolio

narrative profile describing the student's achievements and growth in writing. The organization of the frameworks allowed us to look for consistencies among the features—the extent to which students seemed to be setting goals for themselves, using others' comments, and following through in revision and improvement in subsequent pieces of writing. A sample individual narrative for Barry is provided in Figure 3. In developing and presenting the narratives, we followed Erickson's guidelines of including both *particular descriptions,* to clarify the meaning of a category or conclusion, and *general descriptions,* to display the breadth of evidence that supported a conclusion. Particular descriptions might include a quotation from the student's work, a "narrative vignette" describing the substance or structure of a piece, or a reference to an entire piece of writing attached to the narrative. General descriptions might include tables or prose displaying such information as the number of sequences per genre, the number or percent of sequences containing at least one preliminary draft, the number or percent of sequences in which a student reflected on revision

or word choice (organized by time of year), and so on. Taken together, the frameworks and the narrative leave an evidentiary trail that portfolio readers (parents, colleagues, supervisors, or others) might use to trace the path of evidence leading to a conclusion and, where appropriate, support or question the interpretation.

Validity Concerns and Responses
To investigate the validity of the portfolio-based interpretations, we focused on 10 students in Howard's class, examining both their comprehensive writing folders, containing all writing completed during the year, and their portfolio, containing the pieces they had selected to represent themselves as developing writers. Beck, Ebbs, Herter, Muchmore, and Steele each "adopted" two students and, using the framework and procedures we had developed, wrote extended case studies based on the comprehensive folders and the portfolios. As part of their case studies, they addressed the extent to which the students had adequately represented themselves in selecting pieces for their portfolios. Taylor and Matson independently read each of the 10 portfolios and wrote brief narratives, typical of what a teacher might write, on each student. We were interested in comparisons among the independent readings and between the portfolios and the comprehensive writing folders; we were also interested in the efficiency with which the portfolio-based narratives could be completed.

About the validity of the readings as reflected in the narrative profiles. In traditional terms, our interreader reliability, defined as consistency among independent readings, would be unacceptably low. A content analysis comparing individual profiles for each of the 10 students showed substantial differences in emphases. Some readers focused more on the qualities of the final draft, whereas others focused more on the processes reflected in the revisions and self-assessments; most emphasized the strengths apparent in the portfolios, although one reader tended to emphasize the weaknesses.[1] More discussion and experience with the frameworks and narratives will no doubt lead to higher levels of consistency. However, to aim for high levels of agreement from *independent* readings in this context would miss the point. If this approach is used as intended, the central interpretation will be the classroom teacher's interpretation, and it will be based not only on the portfolios but also on extensive knowledge of the students, their goals, and their instructional opportunities. The process of writing the narratives will leave an evidentiary trail that colleagues, supervisors, and other appropriate stakeholders can follow. The narratives, including particular and general descriptions, taken together with the frameworks and the portfolios, allow another appropriate reader to serve as co-analyst, tracing the evidentiary trail that led to the conclusions and raising alternative interpretations for discussion. These second readings of a sample of portfolios may take place informally in discussions among teachers within a building or a district or more formally in supervisory meetings. Discussions with parents and students might also be integrated with the teacher's interpretations. Taken together, these activities serve the same purpose that multiple independent readings are intended to serve–warranting the validity and fairness of the interpretation. This approach, however, acknowledges the singular value of the teacher's knowledge base in making interpretations, which cannot be duplicated by outside readers. Thus, the validity of the conclusions are warranted in the process of data analysis and the transparency of the evidentiary trail, which allows the reader to trace the teacher's arguments. Differences of opinion between readers are opportunities for discussion and rethinking of initial interpretations as different perspectives are brought to bear.

About the representativeness of portfolio selections. Conclusions about the students, individually and as a class, are of course constrained by the pieces of writing that they chose to share in their portfolios. Some students gave us in-depth information with respect to a particular genre but little information about other genres, while others gave us a broader sampling of writing across genres but insufficient samples to note change within a genre. Most students chose to share earlier drafts, but some did not. Moreover, a careful reading of the folders from which the portfolios were selected showed that students occasionally left what we perceived to be the stronger pieces out of the portfolio. In one sense, this does not undermine the credibility of the portfolio reader's conclusions, because the conclusions can be limited to the available data base, which can be carefully described. In another sense, however, it limits the picture that can be provided of the student as a writer.

It would, of course, be possible to require students to follow specific guidelines in creating their portfolios—to include at least two pieces of writing from at least two genres, for instance. This would result in more representative information; however, it would take some of the ownership of the portfolios away from students and risk undermining the pedagogical and motivational value of the portfolio selection process. Deciding on how to balance students' autonomy with teachers' and others' informational needs will not be easy, and teachers will need to weigh competing priorities in making these decisions.

About the time involved. Perhaps the most obvious concern to raise about the approach we are suggesting is its efficiency. At first glance, it looks like an imposing task. However, with practice, most of us found we could work through a portfolio typically containing six to seven sequences of writing in about half an hour. If the frameworks were completed as each sequence of writing was placed in the portfolios, the end-of-year time commitment could be cut substantially. Moreover, teachers may choose to work with a far less complex list of features as they develop interpretive frameworks relevant to their own instructional contexts. The individual narratives took another half hour although, again, with practice and a better developed system of referencing individual pieces of student writing, that time could be reduced somewhat.

Even if the procedures are streamlined in the ways suggested, the process will clearly add to teachers'

work loads. We would be remiss if we did not also note that schools and districts interested in fostering the professionalization of teaching, in this or other ways, must seriously rethink the way teachers' work loads are structured. Disciplined inquiry and collaboration, both of which are essential to professional practice, require time.

Possibilities for Providing System Level Information

When accountability concerns focus on the progress of individual students or the professional judgments of teachers, then the procedures outlined above for drawing, reporting, and evaluating portfolio-based conclusions are appropriate. At the school level, committees of teachers, supervisors, and others could be formed to engage in periodic audits of the individual portfolios (Adams & Burgess, 1992; Johnston, 1989). At the district and state level, the procedures of these school level committees might be reviewed to ensure that appropriate standards are being followed (Adams & Burgess, 1992). Strategies have been suggested for integrating classroom level information into more comprehensive systems of accountability that also include more standardized measures (e.g., Archbald & Newmann, 1988; Darling-Hammond & Ascher, 1991).

If information from these non-standardized forms of assessment is desired at the program, school, district, or state level, we might try something like the following: We ask teachers to maintain portfolios with all their students—portfolios accompanied by descriptions of instructional goals and activities and by teachers' interpretations of students' work. Periodically, carefully selected random samples of classroom portfolio sets, accompanied by the teacher's commentary, would be collected at the system level. Committees of teachers and researchers would work together to develop a framework, possibly like the one presented above, to describe the contents of the portfolio sets as well as the quality of work contained therein. Categories might address questions like the following: What kinds of projects are students work-

ing on in their classrooms? Do these projects encourage critical or creative thinking? Do they encourage self-reflection and revision? Are the teachers' goals reflected in their students' work? What kinds of feedback are students receiving? How are they using that feedback? What criteria and standards are teachers using to evaluate student work? What strategies are they using to document the validity of their evaluations? This idea of sampling existing classroom work is not new. For instance, the surveys of writing in secondary schools conducted by Applebee in the early 1980s provide useful examples (1981, 1984). This kind of contextualized performance assessment is likely to provide an important supplement to the standardized sorts of performance assessment typically used at the system level and to suggest directions for curricular reform.

Implications

If, as Johnston (1989) and others have suggested, the purpose of assessment is to provide high-quality education for all students and if exclusive adherence to traditional modes and criteria of assessment may be undermining that goal, then we need to rethink our priorities and explore alternative means of meeting accountability needs. Given the well-documented power of assessment to influence what students learn and what teachers teach, it becomes crucial to discover or develop approaches to assessment that reflect the full range of skills, knowledge, and interests that we want to nurture, including, perhaps, the ability to *develop* intellectual problems and evaluative criteria. Interpretive research traditions offer an epistemological foundation and methods well suited to validity research in these radically nonstandardized assessment contexts.

Notes

We are grateful to the Arts PROPEL research team, especially Roberta Camp, Kathryn Howard, and her students, who were generous in sharing their portfolios, writing folders, and insights on their experiences with portfolio assessment. The views expressed in this article are the authors' own. We are also grateful to Loren Barritt, Ruth Behar,

John Friedlander, Anne Ruggles Gere, Deborah Keller-Cohen, Anita Norich, and Stuart Rankin for helpful comments on earlier drafts of this article.

[1]In a detailed comparison of Taylor's profiles with the comments about the portfolios contained in the extended case studies, we found that the percentage of idea units in Taylor's profiles that could be matched ranged from 22% to 70%, with a median of 48%. We also estimated interreader reliability based upon the completed frameworks for each sequence, considering percent agreement and kappa coefficients between the three possible pairs of readers across all categories. For the categories in the final column, percent perfect agreement ranged from 11% to 78%, with a median of 38%; percent agreement within one category ranged from 24% to 95%, with a median of 78%. Also for the final column, kappa coefficients—which correct for agreement due to chance and are substantially reduced when variability is low— ranged from 0% to 48%, with a median of 11% for perfect agreement, and from 0% to 81%, with a median of 30%, for agreement within one category. For the revision, self, and teacher columns, because of the low use of many categories, we looked at agreement by cluster (vision, development, language, style, reader, and writer), asking whether the readers agreed that a revision or comment relevant to that cluster had occurred. (No coefficients are reported for "other," because the portfolios rarely contained others' comments.) Here, across all clusters and columns, percent agreement ranged from 49% to 97%, with a median of 81%, and kappa ranged from 0% to 71%, with a median of 32%.

References

Adams, E., & Burgess, T. (1992). Recognizing achievement. In H. Berlak, F. Newmann, E. Adams, D. A. Archbald, T. Burgess, J. Raven, & T. A. Romberg (Eds.), *Toward a new science of educational testing and assessment* (pp. 117–137). Albany: State University of New York Press.

Applebee, A. N. (1981). *Writing in the secondary school: English and the content areas.* Urbana: National Council of Teachers of English.

Applebee, A. N. (1984). *Contexts for learning to write: Studies of secondary school instruction.* Urbana: National Council of Teachers of English.

Archbald, D. A., & Newmann, F. M. (1988). *Beyond standardized testing: Assessing authentic academic achievement in the secondary school.* Reston, VA: National Association of Secondary School Principals.

Camp, R. (1992). Portfolio reflection in secondary school classrooms. In K. Yancey (Ed.), *In context: Varieties of portfolio practice and assessment* (pp. 61–79). Urbana: National Council of Teachers of English.

Crooks, T. J. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research, 85* (4), 438–481.

Darling-Hammond, L. (1989). Accountability for professional practice. *Teachers College Record, 91*(10), 59–80.

Darling-Hammond, L., & Ascher, C. (1991). *Creating accountability in big city school systems.* New York: Teachers College, Columbia University, National Center for Restructuring Education, Schools, and Teaching. (ERIC Clearinghouse on Urban Education No. 334339)

Erickson, F. (1986). Qualitative methods in research on teaching. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (pp. 119–161). New York: Macmillan.

Goetz, J. P., & LeCompte, M. D. (1984). *Ethnography and qualitative design in educational research.* Orlando: Academic Press.

Holmes Group. (1986). *Tomorrow's schools.* East Lansing, MI: Author.

Howard, K. (1990). Making the portfolio real. *Quarterly of the National Writing Project and the Center for the Study of Writing, 12*(2), 4–7, 27.

Johnston, P. (1989). Constructive evaluation and the improvement of teaching and learning. *Teachers College Record, 90*(4), 509–528.

Johnston, P. (1987). Assessing the process, and the process of assessment, in the language arts. In J. R. Squire (Ed.), *The dynamics of language learning: Research in reading and English* (pp. 335–357). Urbana: ERIC Clearinghouse on Reading and Communication Skills.

Johnston, P. H., Weiss, P., & Afflerbach, P. (1990). *Teachers' evaluation of the teaching and learning in literacy and literature.* Albany: Center for the Learning and Teaching of Literature.

Lincoln, Y., & Guba, E. (1985). *Naturalistic inquiry.* Beverly Hills: Sage Publications.

Madaus, G. F. (1988). Testing and the curriculum: From compliant servant to dictatorial master. In L. Taylor (Ed.), *87th NSSE Yearbook* (pp. 83–121). Chicago: National Society for the Study of Education.

National Board of Professional Teaching Standards. (1990). Request for proposals for the first Assessment Development Laboratory. Detroit, MI: Author.

National Commission on Testing and Public Policy. (1990). *From gatekeeper to gateway: Transforming testing in America.* Chestnut Hill, MA: Boston College, National Commission on Testing and Public Policy.

Newmann, F. M. (1990). Higher order thinking in teaching social studies: A rationale for the assessment of classroom thoughtfulness. *Journal of Curriculum Studies, 22*(1), 41–56.

Resnick, L. B., & Resnick, D. P. (in press). Assessing the thinking curriculum: New tools for educational reform. In B. R. Gifford & M. C. O'Connor (Eds.), *Future assessments: Changing views of aptitude, achievement, and instruction.* Boston: Kluwer.

Smith, M. L. (1991). Put to the test: The effects of external testing on teachers. *Educational Researcher, 20*(5), 8–11.

Willinsky, J. (1990). *The new literacy: Redefining reading and writing in the schools.* New York: Routledge.

Wolf, D. (1989, April). Portfolio assessment: Sampling student work. *Educational Leadership,* pp. 35–39.

Wolf, D., Bixby, J., Glenn, J., III, & Gardner, H. (1991). To use their minds well: Investigating new forms of student assessment. *Review of Research in Education, 17,* 31–74.

CALL FOR AWARD NOMINATIONS

NCME Award for Career Contributions to Educational Measurement

The award is to honor persons whose contributions over a career have had widespread positive impact on the field of educational measurement through (a) theoretical or technical developments, (b) influential ideas or conceptions, (c) unique efforts that have increased public understanding of the field, and/or (d) development of procedures, instruments, or programs of special significance. Nominations are limited to living persons.

Nominations should consist of a 1- to 2-page written statement describing the nature, importance, and impact of the individual's contribution. Mail nominations to Liora P. Schmelkin, Dept. of CRSR, Mason Hall, Rm 212, Hofstra University, Hempstead, NY 11550.

Nominations for the 1993 award must be received by January 4, 1993. The 1993 award will be announced at the Annual Meeting in Atlanta.

NCME Award for Application of Educational Measurement Technology

NCME is pleased to announce its call for nominations for the triennial award for the application of educational measurement technology to a specific problem. This is one of three award competitions held triennially. This year's award will be based on an application completed in 1990, 1991, or 1992. NCME members and others are strongly encouraged to help in identifying candidates for this important award.

Examples of applications to specific problem areas include, but are not limited to, selection or classification of students, measuring a hard-to-measure trait, evaluating an educational program or product, integrating testing and learning, or applying technology in a new way to a current problem. Selection criteria are quality and innovativeness of the application effort and the positive impact of the application on the practice of educational measurement.

Self-nominations are encouraged as are nominations of others, with their permission. Nominees need not be NCME members. A nomination consists of five copies of a 3–5 page statement describing the technology, application area, and products or results. Finalists may be requested to submit additional information.

Nominations should be sent by January 20, 1993, to Linda K. Junker, Department of Research, Evaluation, and Planning, Chicago Public Schools, 1819 W. Pershing Rd., 4W, Chicago, IL 60609 (Fax: 312-535-4168; office: 312-535-4080). The award will be presented at the 1993 NCME Annual Meeting in Atlanta.