DOCUMENT RESUME

ED 350 312 TM 019 007

AUTHOR Baker, Eva L.; Linn, Robert L.
TITLE Writing Portfolios: Potential for Large Scale Assessment. Project 2.4: Design Theory and Psychometrics for Complex Performance Assessment. Design and Analysis of Portfolio and Performance Measures.
INSTITUTION National Center for Research on Evaluation, Standards, and Student Testing, Los Angeles, CA.
SPONS AGENCY Apple Computer, Inc., Cupertino, CA.; Office of Educational Research and Improvement (ED), Washington, DC.
PUB DATE Feb 92
CONTRACT R117G10027
NOTE 28p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Educational Assessment; Elementary Education; *Elementary School Students; Evaluation Methods; Evaluators; Grade 1; Grade 3; Grade 4; *Holistic Approach; Intermediate Grades; *Portfolios (Background Materials); *Scoring; *Student Evaluation; *Writing Evaluation; Writing Improvement
IDENTIFIERS *Large Scale Programs; Performance Based Evaluation

ABSTRACT

The use of portfolio assessment as a method of evaluating the writing competence of elementary school students was studied. The study contained two components: (1) an empirical study of the utility and meaningfulness of using an analytic rubric developed for the evaluation of traditional writing samples to score student portfolios; and (2) a qualitative analysis of scoring approaches, drawing on raters' critiques of the analytic scoring approach. Data came from 5 portfolios for grade 1, 23 portfolios for grade 3, and 6 portfolios for grade 4. The rubric used was a well-motivated and well-researched method. The three raters were experienced in the scoring rubric. Results provide some support for the values of a well-motivated writing rubric both for samples of classroom writing and for portfolio collections. Results indicate that, when compared to traditional writing assessment, holistic ratings of class work and of portfolio collections can be achieved with a high level of rater agreement and the ratings can discriminate among grade level and genre differences in students' competence. Raters raised several concerns about portfolio design, emphasizing that the design of a rubric must be coordinated with the design of the portfolio collection. The importance of considering domains of assessment and the utility of portfolios for large-scale assessment are discussed. One table and 22 references are included. (SLD)

National Center for Research on
Evaluation, Standards, and Student Testing

Final Deliverable – February 1992

Project 2.4: Design Theory and Psychometrics
for Complex Performance Assessment

Design and Analysis
of Portfolio and Performance Measures

Writing Portfolios: Potential for Large-Scale Assessment

Eva L. Baker and Robert L. Linn
Project Directors

U.S. Department of Education
Office of Educational Research and Improvement
Grant No. R117G10027 CFDA Catalog No. 84.117G

Center for the Study of Evaluation
Graduate School of Education
University of California, Los Angeles
Los Angeles, CA 90024-1522
(310) 206-1532

The work reported herein was supported in part under the Educational Research and Development Center Program cooperative agreement R117G10027 and CFDA catalog number 84.117G as administered by the Office of Educational Research and Improvement, U.S. Department of Education.

PREFACE

Eva L. Baker and Robert L. Linn

Among the options for alternative assessment strategies that break away from traditional multiple-choice measures, none has captured the imagination more than portfolio assessment. Extrapolated from "artists'" portfolios, the assessment strategy attempts to provide extended records of student performance, to motivate students to choose and reflect on their efforts, and to provide an occasion for teacher-student classroom instruction, all outcomes in short supply in the typical classroom experience.

There is strong anecdotal and analytical evidence in the work of Sizer (forthcoming), Tierney, Carter and Desai (1991), Wiggins (1989), and Wolf (in press) that portfolios can change the character of instructional interaction and contribute to learning. The challenge with portfolios comes when we attempt to use them as formal measures of educational change. Because of the great enthusiasm that portfolio assessment has engendered, with very few large-scale empirical trials for verification (P. Bergman, personal communication, 1991; E. Roeber, personal communication, 1991), its role in either system or individual student accountability is not clear. Consequently, we believe that a program focusing on the design and analysis of portfolios is necessary.

What criteria should guide portfolio design? Clearly, we would expect portfolios to optimize the validity criteria underlying the CRESST research program (Linn, Baker, & Dunbar, 1992). Portfolio assessment should increase the *meaningfulness* of student performance, provide a context for engagement and an opportunity to review and select new goals. Portfolios could also strengthen, at least in part, the *content quality* of what is assessed, by providing opportunities for deeper analysis of subject matter. Similarly, the *cognitive complexity* of student performance could be strengthened. Where portfolio techniques run into trouble are in the areas of *fairness, transfer and generalizability,* and *cost and efficiency*. Clearly, not every measurement strategy, portfolios included, can fully exhibit each of the criteria of validity.

But assuring fairness, reasonable cost and efficiency are essential if portfolios are to be used in any large-scale or accountability context.

Issues in the Design and Analysis of Portfolios

Among the major issues guiding portfolio design is the intended use of the assessment. If comparisons among children or schools are intended, then some level of comparability of design is requisite. Portfolios that are structured to consist of particular elements provide one start at comparability, although administrative comparability is surely not possible. Another choice point is whether portfolios will be individually scored, and if so, whether the score derives from a judgment of the overall effort—the portfolio in totality—or from the sum of individual pieces, for instance, essays.

We have decided to use an opportunity for the collection of data of portfolios in elementary school to inform our thinking about how portfolios should be designed for given assessment purposes. Our Project 3.2 studying the early implementation of portfolios in a state will provide an additional datapoint later this year. Third, our analysis of portfolios to exhibit workforce readiness skills will provide a third instance. Each of these examples differs in the structure and strategy of portfolio assessment. We believe that our analyses of empirical examples will permit us to provide better guidance to the field than our armchair hopes and desires. To that end, we are providing the results of a research effort, co-sponsored by Apple Computer's Apple Classrooms of Tomorrow^{sm} Project as a data-based example of an effort to structure and score portfolio elements.

WRITING PORTFOLIOS:
POTENTIAL FOR LARGE-SCALE ASSESSMENT[1]

Maryl Gearhart, Joan L. Herman, Eva L. Baker, and Andrea K. Whittaker[2]

Center for the Study of Evaluation

University of California, Los Angeles

In this paper we report an investigation of portfolio assessment as a method of evaluating elementary students' competence in writing. Our study contained two components: (a) an empirical study of the utility and meaningfulness of using an analytic rubric (developed for evaluation of traditional writing samples) to score students' portfolios; and (b) a qualitative analysis of scoring approaches, drawing particularly on raters' critiques of the analytic scoring approach. The analytic rubric used was a well-motivated and well-researched method for writing evaluation and as such offered a solid ground for exploring the scorability of portfolios and for generating possible revisions or additions to the rubric.

---

[1] Our work has received partial support from the Apple Classrooms of Tomorrow[sm] Project, Advanced Development Group, Apple Computer, Inc., and from the Educational Research and Development Center Program, cooperative agreement number R117G10027 and CFDA catalog number 84.117G, as administered by the Office of Educational Research and Improvement, U.S. Department of Education.

The findings and opinions expressed in this report do not reflect the position or policies of Apple Computer, Inc., the Office of Educational Research and Improvement, or the U.S. Department of Education.

[2] Robert Tierney collaborated in the design of the classroom portfolios and methods for their use. Our three raters, Judy Bowers, Bernie Honey, and Karen Perry, made important and substantial contributions that we summarize in detail in the paper. John Novak contributed invaluable computing and statistical advice and assistance. Our thanks to the ACOT teachers, students, and parents who permitted us access to the portfolios.

Background

Large-scale assessment of students' writing samples began in earnest in the 1960s (Freedman, 1991; Huot, 1991). While considered a more direct evaluation of students' writing competences than multiple-choice items, many aspects of the approach have nevertheless been controversial (Dyson & Freedman, 1990; Freedman, 1991; Huot, 1991; Moss et al., 1991). A standard prompt is not likely to reflect equally well the background knowledge or interests of all of the students assessed. The genres and topics tested may not mesh with curriculum, resulting either in a revision of the curriculum to "teach to the test" or an inadequate assessment of what was taught. The testing procedures—including an unfamiliar examiner, a characteristically short time limit (typically 20-40 minutes), and a lack of provision for pre-writing activities and draft revisions—may impair the validity of the assessment results.

Rubrics for scoring students' writing samples have also undergone considerable scrutiny (Freedman, 1991; Huot, 1991). The most common approaches to large-scale writing assessment are: *holistic scoring*, assignment of a single score reflecting a student's competence with all aspects of writing; *analytic scoring*, in which dimensions of good writing are defined that should apply across a range of topics within broadly defined genres; and *primary trait scoring*, with rubrics customized to specific prompts. The distinctiveness among these rubrics has been challenged from time to time by research findings (Freedman, 1991; Huot, 1991), and additional questions, focused on large-scale assessment, address their relative efficiency, cost effectiveness, and value for instructional feedback.

In the context of these debates about appropriate procedures for collecting and rating students' writing samples, there has emerged a growing interest in portfolio assessment as an alternative (Freedman, 1991; Mills, 1989; Murphy & Smith, 1990; Simmons, 1990; Tierney, Carter, & Desai, 1991; Wolf, 1989, in press). Portfolios contain students' class writing assignments, composed and collected under potentially more authentic circumstances than writing "tests." There is greater likelihood that classroom experiences provided students with necessary background knowledge and that the writing was directed at a meaningful purpose for the student (a school newspaper article, a presentation to the class, an opportunity to share with parents). In

addition, in that portfolios consist of collections of work rather than single pieces, they hold promise for a richer and more valid portrait of students' competences and progress.

Approaches to Portfolio Assessment

As a new approach to assessment, "portfolios" have varying meanings (Baker, Gearhart, Herman, Tierney, & Whittaker, 1991; Freedman, 1991; Murphy & Smith, 1990; Simmons, 1990; Tierney et al., 1991; Wolf, 1989). There is variation in the persons who compile and organize the collection, the nature of the contents of the collections, and the functions that the resulting portfolios serve in instruction or assessment.

Several different approaches to scoring of students' portfolio collections are currently in development. For example, analytic rubrics are being trialed in Vermont to assess selected key aspects of competent writing, including sentence variety and sense of personal expression, mechanics, fluency and organization, and skill with draft revision (Mills, 1989; Tierney et al., 1991). Moss et al. (1991) are designing analytic rubrics for summarizing the contents and quality of students' writing on a variety of analytic dimensions (e.g., Vision, Development, Language/Form, Literary Style, Reader's Response, and Sense of Writer); the coded information then serves as a basis for a teacher's narrative evaluation. Lewis (1990) has developed 2-point scales for assessing students' progress along dimensions as varied as composition length, mechanics, and risk-taking, and 5-point scales for effectiveness, growth, and self-direction. Wolf (in press) and collaborating teachers have piloted methods of assessing students' progress along key dimensions such as accomplishment in writing, use of processes for writing, and development as a writer.

While the rubrics being developed vary, these efforts have confronted similar questions regarding portfolio contents and scoring procedures: What kinds of writing samples must be contained in a portfolio to permit reliable and valid judgments? How should these samples be organized? What is the impact on raters' judgments if each piece is scored separately versus scoring the collection as a whole?

In our own work, we have recognized that developing methods of portfolio assessment entails *conjoint* development of *portfolio scoring rubrics* and

criteria for portfolio inclusions. In addition, we have addressed critical questions regarding the quality of the measurement process itself: Can raters reach satisfactory levels of agreement when rating portfolios? Do raters' judgments based on a portfolio collection agree with the average of their judgments of the individual writing samples? Are raters' judgments of students' portfolios a valid assessment of students' competence in writing? Such are the issues of our study.

Our Project

The site for our project has been an elementary school that serves as one of the longitudinal research sites of the Apple Classrooms of Tomorrow^{sm} (ACOT^{sm}) Project. The availability of computer support has been one of several contributors to growing interest in students' writing and to the need for appropriate, well-motivated indices of students' writing growth. In 1989-90, in collaboration with Robert Tierney of Ohio State University, we initiated a pilot design for portfolio assessment (Baker, Gearhart, Herman, Tierney, & Whittaker, 1991). Since then we have been working closely with teachers to explore the potential of portfolios for both classroom and external assessment of student progress in writing.

Our design was modeled on a project in primary school classrooms in Westerville, Ohio (aspects are described in Tierney et al., 1991). The portfolios are composed of both a "working" file and a smaller "showcase" file of students' selections of their best pieces. The teachers provide folders for students' working portfolios and time for students to add to and organize their work; included are all stages of the writing process—prewriting (lists, notes, diagrams, etc.), rough drafts, final drafts, and published pieces—and writing in all curriculum areas. For their showcase portfolios, students periodically select those special pieces that they feel represent their best work—not necessarily the final published versions. The showcase portfolios provide the context for an integrated set of assessment activities: student self-assessment (reflective writing prompted by sentence frames), teacher-student conferencing, informal parent-child conferencing, and parent assessment (responses to several open-ended questions).

In 1989-90 we began with three ACOT teachers representing grades 1, 3 and 4, but the project has grown to include all 15 faculty from grades 1 to 6.

The data reported in this paper are those from our initial 1989-90 classrooms. Material included 34 portfolios (5 for Grade 1, 23 for Grade 3, 6 for Grade 4) sampled across students rated high, medium, and low for writing ability at each grade level. The mean numbers of samples in the portfolios were: Grade 1 (5.4), Grade 3 (9.0), and Grade 4 (4.5). By genre, the mean numbers of samples were: Narratives—Grade 1 (3.8), Grade 3 (4.0), Grade 4 (2.0); Summaries—Grade 1 (0.0), Grade 3 (0.8), Grade 4 (1.3); Other—Grade 1 (1.6, primarily poems), Grade 3 (4.2, primarily poems), Grade 4 (1.2, some letters, some poems).

Contrasts between the ACOT portfolios and traditional writing assessment. The ACOT portfolios represented a departure from traditional writing assessment in three key respects. (a) *Classroom writing*: Portfolio samples were students' classroom writing, rather than responses to a prompt administered under standard conditions. (b) *Multiple samples over time*: The portfolios represented multiple opportunities and a range of contexts for demonstrations of competence collected over time, rather than responses collected at a single administration. (c) *Task variation*: The portfolio samples included different genres and multiple topics within genres. While some large-scale approaches to writing assessment (California Assessment Program, 1989) employ matrix sampling techniques to sample a variety of writing types, the typical approach to student assessment focuses on one or two genres, and for each genre provides students with only one opportunity to respond.

Each portfolio contained samples of a student's writing over approximately a 5- to 6-month period, with samples sequenced by date. Prior to scoring, students' names and grade levels were removed from all material.

Our Study

From a rich range of possibilities, we selected what we considered "first comes first" research questions concerning the scorability of portfolios and the meaning of the resulting scores:

Rater agreement: Can a holistic/analytic scheme be applied to the scoring of classroom samples and to the scoring of portfolio

collections with the same levels of rater agreement typically reported for standard writing assessments?

While critics have raised important questions regarding the validity of standard writing assessments, ratings of such samples by two or more judges do typically show acceptable levels of interrater agreement (Huot, 1991). Could comparable levels of agreement be achieved by raters for classroom assignments that were not highly controlled or standardized? for portfolio collections of classwork?

We selected for trial a well-validated rubric, developed for elementary-level narrative writing, that contained a holistic (General Competence) rating as well as analytic subscales reflecting components of competent narratives (Table 1).[3] The rubric, derived from the same sources as the IEA scales (Gorman, Purves, & Degenhart, 1988), was developed by a southern California school district in collaboration with UCLA, and it is in annual use in district assessments of students' narratives. (See Quellmalz & Burry, 1983, for a description of the original UCLA scales.) The rubric was adapted by our raters for summaries and for portfolio collections (see Gearhart, Herman, Baker, & Whittaker, 1992).[4]

Raters were asked first to rate students' classroom work separated into genre sets. Because narrative and summary were the genres most emphasized across grade levels, we created sets of Narratives and of Summaries which had been copied from students' portfolios; each set was scrambled by date, topic, and grade level. Unknown to the raters, we scrambled into the narrative set students' responses to a Standard Writing Assessment (a narrative) that we had administered in the late spring of 1990. We then asked raters to rate students' portfolio collections; for this task, at the raters' request, grade levels were kept separate and identified. The Portfolio Collections also contained the students' Standard Writing Assessment.

[3] The rubrics for summaries and for portfolio collections are available from the authors in Gearhart, Herman, Baker, and Whittaker (1992).

[4] The rubric had been previously adapted to the scoring of Descriptive and Persuasive writing (Baker, Gearhart, & Herman, 1989, 1991).

Table 1

Elementary Narrative Holistic/Analytic Scale

General Competence	Focus/Organization	Development	Mechanics
6 EXCEPTIONAL ACHIEVEMENT EXCEPTIONAL WRITER	- topic clear - events logical - no digressions - varied transitions - transitions smooth and logical - clear sense of beginning and end	- elements of narrative are well-elaborated (plot, setting, characters) - elaboration even and appropriate - sentence patterns varied and complex - diction appropriate - detail vivid and specific	- one or two minor errors - no major errors
5 COMMENDABLE ACHIEVEMENT COMMENDABLE WRITER	- topic clear - events logical - possible slight digression without significant distraction to reader - most transitions smooth and logical - clear sense of beginning and end	- elements of narrative are well-elaborated - most elaboration is even and appropriate - some varied sentence patterns used - vocabulary appropriate - some details are more vivid or specific than general statements - a few details may lack specificity	- a few minor errors - one or two major errors - no more than 5 combined errors (major and minor) - errors do not cause significant reader confusion
4 ADEQUATE ACHIEVEMENT COMPETENT WRITER	- topic clear - most events are logical - some digression causing slight reader confusion - most transitions are logical but may be repetitive - clear sense of beginning and end	- most elements of narrative are present - some elaboration may be less even and lack depth - some details are vivid or specific although one or two may lack direct relevance - supporting details begin to be more specific than general statements	- a few minor errors - one or two major errors - no more than 5 combined errors (major and minor) - errors do not cause significant reader confusion

Table 1 (continued)

General Competence	Focus/Organization	Development	Mechanics
3 SOME EVIDENCE OF ACHIEVEMENT DEVELOPING WRITER	- topic clear - most events logical - some digression or overelaboration interfering with reader understanding - transitions begin to be used - limited sense of beginning and end	- elements of narrative are not evenly developed, some may be omitted - vocabulary not appropriate at times - some supporting detail may be present	- some minor errors - some major errors - some errors cause reader confusion
2 LIMITED EVIDENCE OF ACHIEVEMENT EMERGING WRITER	- topic may not be clear - few events are logical - may be no attempt to limit topic - much digression or overelaboration with significant interference with reader understanding - few transitions - little sense of beginning or end	- minimal development of elements of narrative - minimal or no detail - detail used is uneven and unclear - simple sentence patterns - very simplistic vocabulary - detail may be irrelevant or confusing	- many minor errors - many major errors - many errors cause reader confusion and interference with understanding
1 MINIMAL EVIDENCE OF ACHIEVEMENT INSUFFICIENT WRITER	- topic is not clear - no clear organizational plan - no attempt to limit topic - much of the paper may be a digression or elaboration - few or no transitions - almost no sense of beginning and end	- no development of narrative elements - no details - incomplete sentence patterns	- many major and minor errors causing reader confusion - difficult to read

The three raters were teachers experienced in using the holistic/analytic rubric for scoring their district's assessments of third- and fifth-grade students' narrative writing competence. Therefore no special training on the rubric was required for narrative samples, although raters began the session by scoring and reaching agreement on a training set of 20 samples and checked agreement midway through each scoring session. Raters worked together to adapt the rubric to summaries, established agreement, and then rated the remaining samples. Raters then followed the same procedure for portfolios, but found that they could apply only the holistic rating to the collections.

We compared interrater agreement for judges' ratings of students' Standard Writing Assessments, classroom work presented as sets of Narratives or of Summaries, and Portfolio Collections. We examined both exact agreement and agreement within one score point (the conventional index). While overall agreement across type of assessment was satisfactory, it was highest for Portfolio Collections (.97 to 1.00 for exact agreement); agreements for the other genres varied (for exact agreement, range from .16 to .63; for agreement ±1, range from .80 to .99). Differences in agreement among the subscales were inconsistent across assessment types: Thus, for some subscales, agreements were better for one assessment type (i.e., Standard Assessment, Classroom Narratives, Classroom Summaries, or Portfolio Collections), while for other subscales, agreements were better for a different assessment type. There were no consistent differences among rater pairs in levels of agreement.

To examine the stability of the scheme across raters and rating occasions, we made comparisons of the Standard Writing Assessment scores assigned by these raters with those assigned by a prior group of raters (who had rated ACOT writing samples from five different ACOT schools and their respective comparison groups in the summer of 1990). Since we were comparing the average of the 1990 ratings with the average of the 1989 ratings, we examined agreement ±0.5 and ±1.0. Pointing to scale stability over time, there was very high agreement between this group of raters and the earlier group: for ±0.5 (.76 to .79); for ±1.0 (.94 to .97).

What evidence indicates that raters' judgments of students' portfolios are valid?

We inferred validity from several findings:

- *Grade level differences:* Scores were generally greater for higher grade levels. The General Competence scores for each genre were: Classroom Narratives—2.70 at Grade 1, 3.05 at Grade 3, and 3.39 at Grade 4; Classroom Summaries—2.74 at Grade 3 and 4.36 at Grade 4; Portfolio Collections—3.20 at Grade 1, 3.42 at Grade 3, and 4.00 at Grade 4; Standard Assessment—2.59 at Grade 3 and 3.64 at Grade 4.

- *Genre differences:* Scores were typically higher for the genre emphasized at each grade level. At Grade 3, the Narrative mean was 3.05 compared with 2.74 for Summary; in contrast, at Grade 4, the Narrative mean was 3.39 compared with 4.36 for Summary.

- *Relationships of scores across types of assessments:* The scores for classroom writing (Narratives, Summaries, and Portfolio Collections) were generally similar to one another, and different from the scores for the writing "test" (Standard Writing Assessment); the scores for the narratives (Narratives, Standard Assessment) were similar to one another.

- *Raters' confidence in their portfolio judgments:* In post-rating discussions, raters expressed confidence in their ratings of portfolio collections.

Are portfolio scores based on an aggregate of individual sample scores comparable to raters' single judgments of portfolio collections?

Narratives and Summaries constituted the bulk of the scorable samples in the portfolios (most of the remaining samples were poems that were diverse in form and typically very brief). Therefore, as an alternative portfolio index, we computed the mean of students' individual Classroom Narrative and Summary scores, and we compared this index with the raters' single holistic judgments of Portfolio Collections.

The results indicated that judgments based on individual samples and holistic judgments of collections were typically quite similar, but that, when they differed, holistic judgments of collections were typically higher. It appeared that raters tended to make their whole portfolio judgments based on the more competent pieces in the collections.

Consistency of students' performance across writing contexts: Did students perform comparably in the classroom and in the standard assessment?

Effects of assessment type on raters' judgments: What was the relationship between type of rating material and raters' judgments of students' competence?

We pair these questions together because our design did not permit us to untangle cleanly the effects of task on students' performance from the effects of assessment type on raters' judgments. We employed several statistical approaches to data analysis, and the results are described in detail in Gearhart et al. (1992).

The results suggested that:

- Students were fairly consistent in their abilities to organize their writing across task contexts that differed markedly in genre, topic, and length.
- Differences in task context (e.g., in the classroom—students' access to resources, time, and the assistance of others) had greater impact on the extent and quality of the compositions' development and mechanics than on focus and organization.
- Students' classroom compositions tended to be judged of higher quality than their standard writing assessment samples. Since raters were unaware that the standard assessments (mixed in with the classroom narratives) were not classroom work, the results suggest an effect of task and context on students' performance. However, additional factors implicated include rater bias (the standard assessment responses were briefer and lacking illustrations) and instructional support (students' class work is assisted by others).
- Students' whole portfolios were sometimes rated more highly than their individual samples of classroom work, suggesting an effect of rating task on raters' judgments.

What are raters' opinions of the utility of an analytic rubric for portfolio scoring?

As experienced raters of traditional writing samples and as experienced elementary teachers developing methods of portfolio assessment in their own classrooms, our judges were able to use their expertise to critique the appropriateness of the analytic rubric for portfolio scoring and to suggest

alternatives. The raters' focus group discussions about their scoring experience centered on one central issue: *Analytic rubrics have potential, but portfolio contents need to be structured to provide scorable information.*

The Need to Structure the Portfolio Contents

Mix of genres and topics. The portfolios reflected the considerable mix of genres and topics assigned to students. The raters felt that the mix of genres obscured evidence of the components of writing competence (e.g., organization, style, mechanics) and of change over time in writing quality. As a result, raters were able to assign only a General Competence score. Comparing an October folk tale with a December fantasy, a January haiku, a March whale report, a May letter to a penpal, and a June summary of a field trip was an impossible task. Additional concerns were raised that *task difficulty* and *task familiarity* may have varied unsystematically over time and across students: Students might have more experience with a particular genre, or more background knowledge for certain topics.

Sampling. The portfolios we provided varied markedly in the number of pieces included. The variation reflected (erratically) the number of writing opportunities provided, the number of assignments completed, and/or the number of assignments that students remembered to put into their portfolios. Our raters agreed that, below some minimum number of samples (perhaps six), there was insufficient material to judge overall quality. Number also led to questions about curriculum (the writing opportunities provided), about students (amount of writing undertaken and investment in compiling a portfolio), and about procedures for choosing portfolio samples (especially, student vs. teacher choice). The raters were not certain that students had had sufficient opportunity either to write and/or to complete their portfolios. They worried that students rather than teachers had made decisions about which pieces to include, because they felt that teachers would have a better understanding of how writing reflects competence. Thus, in general they were uncertain that the portfolios were adequate samples of students' work.

Whose Work is Being Assessed?

The need for information on the contributions of others to students' work. For standard writing assessments, students compose their responses

independently. In contrast, students' schoolwork is almost always assisted in some way by teachers, peers, or parents. Although our raters were strong advocates for portfolio assessment, they nevertheless raised questions regarding the validity of writing assessments based on teacher- or other-assisted classroom samples, particul rly if the support and assistance of others varied unsystematically across samples. However, they were not happy with the alternative of portfolio structures with prescribed assignments written under prescribed conditions. What emerged were evident conflicts between their roles as teachers and as raters, between their interpretations of an ideal portfolio for classroom use and a scorable portfolio for external evaluation.

Raters' beliefs about the contributions of word processing to students' work. To date, students' responses to most traditional writing assessments are handwritten, particularly at the elementary level. Having never evaluated word-processed writing samples and not using computers in their own classrooms, our raters believed incorrectly that spellcheckers automatically correct spelling and worried that the Mechanics score was artificially inflated. (Whether any of the raters "adjusted" her scores cannot be determined.) They were concerned that the help of others was "hidden" in word-processed text in ways less likely in handwritten text, even though they were told that all samples were final drafts. They also perceived many word-processed samples as above average in length but not necessarily in quality, and reported irritation at stories that went "on and on and on." Thus, raters may have beliefs about word-processed text that could affect their judgments about students' writing competence. Since word processing can indeed serve different functions in writing (e.g., ongoing use through all phases of writing, typing of final drafts only, use of a spellchecker), raters should have both a general understanding of the functions of word processors and specific information regarding the computational support used for a given piece.

The Raters' Need to Understand Teachers' Expectations

Need for assignment description. In standard writing assessment, raters are informed of the prompts administered to students and adapt the rubric by establishing prompt-specific criteria for each score point. Our raters were accustomed to this procedure and believed that the lack of documentation of

students' assignments impaired their ability to judge the quality of the products. However, since raters' agreement was generally acceptable despite their discomfort, we cannot be certain that their judgments were in fact impaired. How knowledge of an assignment (and other task information, such as a teacher's expectations for the product) may impact raters' judgments is an empirical question.

Mixed grade sets—A need for grade level benchmarks? At the raters' request, we separated and identified the portfolios by grade level. Their request was a result of problems they perceived when applying the rubric to mixed grade samples in the prior rating session: They had never encountered mixed grade samples, since their school district arranges for the scoring of writing samples at separate sittings for each grade level, and grade level is identified. The raters' discomfort with unknown grade levels was interesting, since analytic rubrics can be applied independent of grade-specific competence. Our discovery that they had constructed—but had not formalized—differentiated, grade-specific criteria for assigning scores raised issues about the need for elevating the implicit to the explicit: Should analytic schemes be adapted to assess students' competence in achieving grade-level benchmarks?

Matching Design and Purpose

Conflicts between concepts of "portfolios" and the design requirements for portfolio assessment. It was interesting to hear what our raters thought they would find in the portfolio collections. First, they viewed writing as deeply integrated with language arts and found the limitation to writing somewhat artificial. As elementary level teachers, they were experimenting with language arts portfolios that were far broader in scope in their own classrooms: Their students included in their portfolios audiotapes of oral reading, videotapes of class presentations and performances, logs of books read, and journals, as well as writing. Second, the raters felt constrained by the exclusion of pre-writing and early drafts, because they were deeply engaged in teaching writing as a process, and in using "writing to learn"—about writing, about language use, about books read or experiences. Third, they regarded a portfolio as very much a student's construction and expected to find reflective writing—students' self-assessments and commentary on their feelings about writing, their growth in writing, the value of writing.

We did not ask the raters to evaluate students' competences with language across a range of media, their abilities to plan and revise their compositions, or their understandings of their strengths and weaknesses. If we had, then of course the material the raters felt was missing would have been necessary inclusions. But their concerns raise important issues concerning what constitutes a portfolio: It is clear that conceptions of portfolios are not currently clearly articulated with models of their use for assessment.

The purpose of portfolio assessment. As teachers, our raters were concerned to see that results of assessments serve to guide instruction and its goals. From this standpoint, they suggested teacher-friendly revisions of the holistic/analytic scheme and supplementary assessment dimensions. First, as revisions, they suggested adaptations of the rubric that would enable a teacher to make "commendations" on achievement and "recommendations" for needed improvements. For example, for Organization, a commendation might be "subject clear"; a recommendation might be "include a beginning, middle, and end." The impact of such a revision would be to discourage teachers' use of portfolio assessment solely for summative evaluation and, instead, encourage its use for formative evaluation and redesign of instruction.

Second, our raters felt that, even if a portfolio structure could be designed to support the scoring of portfolio collections with all subscales of an analytic rubric, such a rubric still would fall short of capturing additional scorable dimensions of students' writing competence and students' attitudes toward writing. Potential dimensions suggested included: creativity, perseverance or investment, excitement or interest, openness or willingness to share feelings and ideas, and risk-taking or willingness to try difficult assignments or new forms of writing even if the product is not of acceptable quality.

Summary and Discussion

The purpose of our study was to examine the feasibility of evaluating students' writing competence with an analytic rating of their portfolio collections. Our results provided some support for the value of a well-motivated writing rubric both for samples of classroom writing and for portfolio collections. Results demonstrated that, when compared to traditional

writing assessment, holistic ratings of class work and of portfolio collections can be achieved with high levels of rater agreement, and the ratings can discriminate among grade level and genre differences in students' competence. Ratings of portfolio collections were particularly high, suggesting that the multiple samples contained within a portfolio provide a more comprehensive basis for judging writing quality and thereby support uniformity of judgment. However, our additional results indicating that raters sometimes rate collections higher than the average of their ratings of single pieces suggests something more complex—that a collection may provide a context for anchoring judgments of the better pieces in the collection.

The generally satisfactory levels of agreement are particularly noteworthy in the context of our raters' perceptions of the difficulty of our unconventional procedures. Our raters were not comfortable rating the classroom material without knowledge of the assignment or of students' grade levels; they also found the mix of assignments confusing. As a result, they worked very slowly. Nevertheless, they reported confidence in their judgments, and it appeared that the analytic scheme provided criteria for scoring that were interpreted in a consistent manner across raters, writing assignments, genres, and samples versus whole portfolios.

Thus, the portfolio ratings demonstrated properties that support the utility of at least a holistic portfolio score for writing evaluation. Nevertheless, other results raised issues about the meaning of our portfolio scores. Comparisons of ratings across type of assessment indicated that raters may make somewhat different judgments of students' writing competence depending on the type of assessment. Of key importance was the finding that judgments of students' writing competence may differ when based on portfolio collections rather than responses to standard writing assessments: specifically, raters may score students' competence higher based on portfolio collections, and portfolio scores based on holistic judgments may be higher than those based on aggregates of individually scored samples. Our results raised issues regarding the meanings of portfolio scores achieved through differing rating procedures and aggregated through differing statistical procedures.

In focus groups, our raters raised provocative issues regarding the design of portfolio assessment. (Some of their issues were also raised in focus groups

reported by Meyer, Schuman, & Angello, 1990.) As teachers engaged in portfolio use in their own classrooms, they were hopeful that portfolio assessment can offer a means of evaluation that is more valid than traditional writing assessment. They felt that an analytic rubric has potential for portfolio assessment—provided the subscales reflect teachers' objectives for their students' growth and competence. They raised a number of concerns about the scorability of portfolios. The contents of portfolios need to be structured to suit the purposes of the assessment. There must be some way to provide raters information regarding teachers' expectations for students' performance—for example, description of the tasks assigned to students and of the "benchmarks" used to evaluate competent writing performance at each grade level. Raters need understanding of the students' unique contributions to the portfolio samples: How much assistance was provided by others, by the computer?

Designing a Portfolio Structure for Assessment

Our results are based on the assessment of just one approach to the design of a scorable portfolio—a collection of students' final products sequenced by date—and just one rubric—a holistic/analytic scheme. Given the state of the art in alternative assessment, our approach represented a reasonable first step, and the work has raised a number of critical issues regarding portfolio assessment as an approach to the evaluation of students' competence in writing.

The design of a rubric must be coordinated with the design of a portfolio collection. Portfolios should be displays of work that teachers (and students) believe reveals students' competence along dimensions assessed by raters and known and understood by teachers. The portfolios that we presented were not constructed with those purposes in mind, and therefore it is not surprising that raters were able to assign no more than a holistic score.

Two issues merit special attention in designing a scorable portfolio (cf. Meyer et al., 1990). The first has to do with the selection of separable *domains for assessment* that can set the criteria for portfolio inclusions. There is ample evidence, both from our raters' discussions and from interviews with the teachers participating in the portfolio project, that teachers have difficulty defining domains or separating students' work into domains. Their difficulty is just as likely to be borne of sophisticated curriculum knowledge as

ignorance. Teachers quite knowledgeable about current "whole language" approaches, for example, may conceive of competences as deeply integrated with one another, so that separating domains for purposes of assessment then appears to violate their objectives for their students. Unfortunately, these kinds of conceptions do not support the design of "assessable" portfolios.

A second issue involves the tension between portfolio structures useful for large-scale assessment and those useful as supports for classroom instruction. Our raters' enthusiasm for portfolios reflected the hopes of many teachers that portfolio contents can reflect the full range and depth of their students' activities throughout the year. Yet utility for large-scale assessment requires comparability of portfolios across classrooms and portfolio contents to support credible assessments. The comparability and valid inference requirements necessitate prestructuring of portfolios, which may interfere with teachers' instructional practices. Indeed, a "top-down" portfolio structure could negate some of the "bottom-up" appeal of portfolio use to teachers. Needed are strategies that balance the tension between evaluators' needs to constrain and structure portfolios for assessment and teachers' needs to devise portfolio uses that ensure their discretion in curriculum.

How can we accommodate assessment needs in the curriculum? Possibilities may include: "mini-portfolios" for particular writing projects, collection of multiple samples for each genre during the year to track progress within genre, or establishment of grade level "benchmarks" for writing quality. Any of these would require reorganization of the curriculum, but teachers might find some less restrictive than others. Whatever the solution, it is clear that no set of criteria for a teacher-selected portfolio for external evaluation can be developed without a coordinated framework articulating relationships between curriculum and assessment design.

Our study has confronted us with the complexities entailed in developing methods of large-scale portfolio assessment that can provide useful information about students' competences to teachers, students, parents, and policy makers. We have noted conflicts among practitioners' concepts of portfolio collections and the need for constraints on those collections if they are to be used for large-scale assessment. Required are methods of portfolio use that inform teachers' curriculum and instruction without limiting them, that permit student construction and participation, and yet that are sufficiently

uniform in structure and content to make possible meaningful comparisons among students. It is almost certain that there is no single solution to the multiple functions being advocated for portfolios in and out of the classroom. The challenge is to design multiple prototypes suited to the diverse needs of schools.

References

Baker, E.L., Gearhart, M., & Herman, J.L. (1989). *Apple Classrooms of Tomorrow^{sm}: The 1988-1989 evaluation report.* Los Angeles: University of California, Center for Technology Assessment/Center for the Study of Evaluation.

Baker, E.L., Gearhart, M., & Herman, J.L. (1991). *Apple Classrooms of Tomorrow^{sm}: The 1989-1990 evaluation report.* Los Angeles: University of California, Center for Technology Assessment/Center for the Study of Evaluation.

Baker, E.L., Gearhart, M., Herman, J.L., Tierney, R., & Whittaker, A.K. (1991). Stevens Creek portfolio project: Writing assessment in the technology classroom. *Portfolio News*, *2*(3), 7-9.

California Assessment Program. (1989). Sacramento: California State Department of Education.

Dyson, A.H., & Freedman, S.W. (July, 1990). *On teaching writing: A review of the literature* (Occasional Paper No. 20). Berkeley: University of California, Center for the Study of Writing.

Freedman, S. (May, 1991). *Evaluating writing: Linking large-scale assessment testing and classroom assessment* (Occasional Paper No. 27). Berkeley: University of California, Center for the Study of Writing.

Gearhart, M., Herman, J.L., Baker, E.L., & Whittaker, A.K. (1992). *Writing portfolios at the elementary level: A study of methods for writing assessment* (CSE Tech. Rep. No. 337). Los Angeles: University of California, Center for the Study of Evaluation.

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

Huot, B. (1991). The literature of direct writing assessment: Major concerns and prevailing trends. *Review of Educational Research*, *60*, 237-263.

Lewis, L. (1990). *Pilot project for portfolio assessment.* Fort Worth, TX: Fort Worth Independent School District, Keystone Writing Project.

Linn, R.L., Baker, E.L., & Dunbar, S.B. (1992). Complex, performance-based assessment: Expectations and validation criteria. *Evaluation Comment*, Winter, 2-9.

Meyer, C., Schuman, S., & Angello, N. (September, 1990). *Aggregating portfolio data* (White Paper). Lake Oswego, OR: Northwest Evaluation Association.

Mills, R.P. (1989, December). Portfolios capture a rich array of student performance. *The School Administrator*, 8-11.

Moss, P.A., Beck, J.S., Ebbs, C., Herter, R., Matson, B., Muchmore, J., Steele, D., & Taylor, C. (1991). *Further enlarging the assessment dialogue: Using portfolios to communicate beyond the classroom.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, April.

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

Quellmalz, E., & Burry, J. (1983). *Analytic scales for assessing students' expository and narrative writing skills* (Resource Paper No. 5). Los Angeles: University of California, Center for the Study of Evaluation.

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

Sizer, T.R. (forthcoming). *Horace's school: Redesigning the American high school.* Boston: Houghton-Mifflin.

Tierney, R.J., Carter, M.A., & Desai, L.E. (1991). *Portfolio assessment in the reading-writing classroom.* Norwood, MA: Christopher Gordon.

Wiggins, G. (1989). A true test: Toward more authentic and equitable assessment. *Phi Delta Kappan, 70*, 703-713.

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

Wolf, D.P. (in press). Assessment as an episode of learning. In R. Bennett & W. Ward (Eds.), *Construction versus choice in cognitive measurement.* Hillsdale, NJ: Erlbaum.