CRESST

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

Technical Report

You can view this document on your screen or print a copy.

- UCLA Center for the Study of Evaluation

in collaboration with:

- University of Colorado
- NORC, University of Chicago
- LRDC, University of Pittsburgh
- The RAND Corporation

Writing Portfolios at the Elementary Level:
A Study of Methods for Writing Assessment

CSE Technical Report 337

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

National Center for the Study of Evaluation,
Standards, and Student Testing (CRESST)
UCLA Graduate School of Education
University of California, Los Angeles
Los Angeles, CA 90024-1522
(310) 206-1532

The work reported herein 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, the Office of Educational Research and Improvement, or the U.S. Department of Education.

Acknowledgement

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.

WRITING PORTFOLIOS AT THE ELEMENTARY LEVEL:
A STUDY OF METHODS FOR WRITING ASSESSMENT

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

The need for alternatives to standardized testing in all subject areas has become a national concern. Many efforts are under way to design assessments that better represent competences required for “authentic” and valued work in our society, and that have the potential to provide a more equitable and a more sensitive portrait of students’ strengths and weaknesses. One approach to assessment of writing that holds promise for meeting these objectives is the writing portfolio. Unlike traditional writing assessments based on students’ timed responses to a standard prompt, portfolios can contain writing projects that have engaged students in purposeful writing—a process of background readings and experiences, pre-planning, and opportunities for revision.

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 a holistic/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 basis for exploring the scorability of portfolios and for generating possible revisions or additions to the rubric.

Background: Approaches to Writing Assessment

Large-scale assessment of students’ writing based on writing samples (rather than multiple-choice items) began in earnest in the 1960s (Freedman, 1991; Huot, 1991). While considered a more direct evaluation of students’ writing competences than multiple-choice items, the approach has nevertheless been controversial (Dyson & Freedman, 1990; Freedman, 1991; Huot, 1991; Moss et al. 1991). Serious questions have been raised about the appropriateness of standard writing tasks for students who vary in background knowledge and/or topic interest; such variability may influence

performance regardless of writing ability. Concerns also have been aired that genres and topics tested may not mesh with curriculum, resulting either in "teaching to the test" or an inadequate assessment of what was taught. Finally, writing assessment procedures have been criticized for their white-coat, laboratory approach (test administrators unknown to the students), for the characteristically short time limit (typically 20-50 minutes), and for the lack of provision for pre-writing activities and draft revisions.

Rubrics for scoring students' writing have also undergone considerable scrutiny and revision (Freedman, 1991; Huot, 1991). A common approach to large-scale writing assessment is holistic scoring—assignment of a single score reflecting a student's competence with all aspects of writing. A second approach is primary trait scoring, where scoring rubrics are customized to specific prompts. A third method is the analytic rubric, in which dimensions of good writing are defined that should apply across a range of topics within broadly defined genres. Debates among these approaches focus on their efficiency, cost effectiveness, and relative value for instructional feedback. However, empirical comparisons frequently show significant correlations among analytic subscale scores or across the different types of scoring. (These patterns suggest that the different scoring approaches are not as consistently differentiable as advocates might like to believe.)

In the context of debates about appropriate procedures for collecting and rating students' writing samples, there has emerged a growing interest in portfolio assessment as an alternative that may transcend some of the concerns (Freedman, 1991; Mills, 1989; Murphy & Smith, 1990; Tierney, Carter, & Desai, 1991; Wolf, 1989, in press). Portfolios contain student writing composed under what can be more authentic circumstances than writing "tests." Because the portfolios grow out of classroom assignments, there is greater likelihood that students' writing represents shared experiences with common resources (background knowledge) and that the work accomplishes a meaningful purpose for the student (school newspaper, presentation to the class, sharing with parents). Furthermore, in that evaluators rate collections of work rather than single pieces, portfolios hold promise for a richer and more valid portrait of a student's competences.

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; 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. When portfolios are used for student assessment, assessment practices differ in their purposes (ranging from global “celebrations” of students’ accomplishments to summative grading progress), the persons involved (teacher, outside evaluator, parent, peer, student him/herself), and the form of the assessment (grades, holistic or analytic scoring, narrative commentary).

Several different approaches to scoring students’ portfolio collections are currently in development. For example, analytic scales are being trialed in Vermont as a method of assessing 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 rubrics for representing 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—with each dimension containing from 4 to 9 subcategories). 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 (defined for each grade level), 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 (e.g., awareness of the needs of the audience; organization and development), use of processes for writing (e.g., prewriting and draft revision), and development as a writer (e.g., use of writing for different purposes). The National Assessment of Educational Progress (NAEP) has conducted a pilot evaluation of portfolios focusing on available essays in narrative, informative, and persuasive genre. The scoring guides focus on the degree of development of writing in each of the genre (Gentile, forthcoming).

While the rubrics being developed vary, the efforts have confronted similar questions regarding portfolio contents: 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? These are the issues of our study.

Our Project

The site for our project has been an elementary school that serves as a longitudinal research site for the national Apple Classrooms of Tomorrow (ACOT^{SM}) project. The availability of computer support has been one of several contributors to the school's growing interest in students' writing and to ACOT's desire 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 et al., 1991). Since then we have worked closely with teachers in exploring the potential of portfolios for both classroom and external assessment of student progress in writing.

Our design was modeled on a project we observed 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, 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). The conferences are used to affirm students' strengths and build an atmosphere for writing as a meaningful and informative source of student growth.

In 1989-90 we began with three ACOT teachers representing Grades 1, 3 and 4, but the project grew rapidly in 1990-91 to engage most of the school faculty (six ACOT teachers and four non-ACOT teachers) from Grades 1 to 6. The data reported in this paper are those from our initial 1989-90 classrooms, and all descriptions below pertain to the project as it unfolded during our first year.

Contrasts Between the ACOT Portfolios and Traditional Writing Assessment

The ACOT portfolios represented a departure from traditional writing assessment in five 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. (d) *Writing process:* Many of the portfolio samples had undergone repeated revision. While some traditional writing assessments also provide opportunity for pre-writing activities and revision, many do not, and the time available for any revision, when it is permitted, is limited. In our portfolios, furthermore, the final drafts were attached to pre-writing and early drafts, providing a potential opportunity to assess the process of composing as well as the quality of the product. (e) *Supplemental materials*: The portfolios contained materials in addition to students' writing: students' showcase

selections of their best work and their self-assessments of these pieces, notes from teacher-student conferences, and parents' assessments. These materials could provide additional information on students' knowledge about writing, their assessments of their writing competence, and their attitudes toward writing that is not available from writing samples alone.

Our Study

Clearly, compared with a single writing sample, these portfolios contained a rich diversity of material. Their intuitive appeal for assessment was the quantity, diversity, and range of their contents. From a great 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 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 when writing tasks were not highly controlled nor standardized?

We selected for trial a well-validated analytic rubric that contained subscales reflecting key aspects of competent writing in each of the genres emphasized at our ACOT site. We compared interrater agreement for judges' ratings of (a) students' standard writing assessments, (b) students' portfolio collections, and (c) students' classroom work presented as sets of narrative or descriptive writing, unidentified by student and scrambled by date, topic, and grade level. For this study, portfolio collections were restricted to final drafts of all assignments sequenced by date.

We also made comparisons of the standard writing assessment scores assigned by these raters with the scores assigned to the same responses by a prior group of raters (who had rated ACOT writing samples from five different ACOT schools in the summer of 1990). The results provided an index of the stability of the scheme across raters and rating contexts.

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

We inferred validity from grade level differences (scores should increase with age), from relationships of scores across types of assessments (e.g., scores for portfolio collections and for classroom material should be correlated), and from raters' confidence in their portfolio judgments (based on opinions expressed in post-rating discussions).

For these and most of the subsequent analyses, we constructed two contrasting indices of the competence of students' portfolio writing: the raters' single judgments of a portfolio collection versus the mean of their aggregated judgments of separate narrative and descriptive samples. Differences between the patterns of results for these two indices revealed effects of these assessment types on raters' judgments.

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 paired these questions 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. Thus, for example, a student's classroom work may receive a higher score than her response to a standard writing task because the classroom writing is of higher quality; in the classroom, she had access to resources and had time to revise her composition.[1] On the other hand, a rater might be biased and assign a higher score to a sample from a student's classroom work, because, unlike the standard response, the classroom piece is accompanied with illustrations, is fairly lengthy, and has a jazzy title.

Using several different statistical procedures, we examined interrelationships among scores across the three types of assessment. Because our raters were experienced and were applying a well-validated rubric, we had reason to believe that they would ignore such irrelevant variables as variations in composition length or inclusion of illustrations when scoring the students' writing. Nevertheless, we acknowledge that raters could

[1] If a student's classroom work is of higher quality because she received assistance with it, however, the higher score does not necessarily indicate that she performed more competently in the classroom than on the "test." Instructional support in the classroom creates problems for large-scale evaluation of students' classroom writing, as we discuss later in this paper.

have been influenced by characteristics of material, and we consider rater bias when interpreting results.

What are raters' opinions of the utility of a holistic/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.

Procedures

Datasets

Our portfolio work yielded samples of portfolios from Grades 1, 3, and 4. All materials were labeled with student identification numbers and the date; students' names and grade levels were removed.

For *scoring individual classroom samples,* we first categorized assignments by genre, producing two sets of narratives and summaries, and eliminating all other pieces from these sets. (Remaining were samples of poetry and a few letters.) The narratives and summaries were presented as two separate sets of writing samples (narratives, summaries) each scrambled by grade, assignment type, and date. Unknown to the raters, we scrambled within the narrative set the narratives that were the third- and fourth-grade students' responses to a standard writing prompt ("A Very Special Memory," Appendix A) that had been administered in the late spring of 1990 concurrent with the portfolio work.

For *scoring portfolio collections,* we removed from the portfolios all prewriting, early drafts, student self-assessments, teacher conference notes, and parent assessments, retaining the final drafts of all assignments. We added the standard writing assessment response, and then sequenced all material by date.[2] Table 1 describes the resulting datasets. As we will discuss, there were marked differences among the grades in the number and genres of pieces in the collections.

[2] As we discuss later, our inclusion of the standard assessment somewhat muddies our interpretation of the portfolio scores. Our objective at the time was to have the raters use the sequenced material as a basis for assessing progress over time—including the relation of the quality of the standard response to the quality of prior and subsequent classroom writing samples.

Table 1

Contents of Portfolio Collections

		Classroom Narratives			Classroom Summaries			Other			Total[a]		
Grade	*N*	Mean	*SD*	Range	Mean	*SD*	Range	Mean	*SD*	Range	Mean	*SD*	Range
1	(5)	3.80	.45	3-4	—	—	—	1.60	1.82	0-4	5.40	2.07	3-8
3	(23)	3.96	2.03	1-9	.83	.89	0-3	4.22	2.52	0-10	9.00	4.73	2-20
4	(6)	2.00	1.41	0-4	1.33	.82	1-3	1.17	1.98	0-2	4.50	2.43	1-7

[a] This total represents the total number of classroom samples. In addition, the responses of those students who completed the Standard Writing Assessment were contained in the portfolios presented to raters.

Scoring Rubric

Because we were engaged in supporting classroom as well as large-scale assessment of portfolios, a holistic/analytic rubric that could provide direct feedback for instruction was deemed the appropriate choice. The rubric used to assess all material was a holistic/analytic scheme developed for elementary level writing by a southern California school district in collaboration with the UCLA Center for the Study of Evaluation. (See Quellmalz & Burry, 1983, for description of original CSE scales.) We have used this rubric in our evaluations of ACOT elementary students' writing for the past three years, and it is in annual use in assessments of students' narratives in the Conejo Valley (CA) school district. Originally developed just for narratives, the rubric had also been adapted for descriptive and persuasive writing, and for this study was adapted for summaries and for portfolio collections. Although the rubric contains four 6-point scales for assessing General Competence, Focus/Organization, Elaboration, and Mechanics (Tables 2 and 3), raters restricted their adaptation of the rubric for whole portfolio scoring to the General Competence score (Table 4).

Finally, we asked the raters to trial a scheme for assessing students' progress based on the sequenced material in the portfolio collections (Table 5). For the majority of portfolios, the raters felt unable to assign a score. Their difficulties are summarized in a later section.

Table 2

Elementary Narrative 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 pattern 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 2 (continued)

General Competence	Focus/Organization	Development	Mechanics
3 SOME EVIDENCE OF ACHIEVEMENT DEVELOPING WRITER	- topic clear - most events logical - some digression or over-elaboration 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 - no fewer than 5 combined errors (major and minor) - 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 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

Table 3

Elementary Descriptive/Summary Analytic Scale

General Impression	Focus/Organization	Concrete Language	Elaboration	Mechanics
6 EXCEPTIONAL ACHIEVEMENT EXCEPTIONAL WRITER	- controlling idea clearly stated, unifying and focusing the writing - exceptionally consistent attitude toward subject - effectively orients reader to subject - provides reader with strong sense of closure on the subject	- writer uses specific, concrete language to help reader visualize, hear, feel, smell and taste - all details are consistent with overall intent of writer - details create clear, vivid images - concrete language used realistically or metaphorically to develop the description and its context	- extended elaboration of one main point (usually 8-10 clauses or more)	- one or two minor errors - no major errors
5 COMMENDABLE ACHIEVEMENT COMMENDABLE WRITER	- controlling idea stated or easily implied to provide focus for writing - provides generally consistent attitude toward subject - well organized and unified according to a definite plan - effectively orients reader to subject - provides reader with definite sense of closure on the subject	- writer uses specific, sensory details - most details consistent with intent of writer - metaphorical language, when present, may sometimes seem trite or inappropriate	- full elaboration of one main point (usually 6-9 clauses)	- 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	- controlling idea present, but may not provide a definite focus for writing - controlling idea may allow for some inconsistency of attitude toward subject - plan is present but writer may occasionally deviate from the plan - reader is oriented to the subject - may end awkwardly or abruptly	- writer uses some specific, sensory detail - most details are consistent with overall intent of writer - details generally create clear images - where used, concrete language is used realistically	- moderate elaboration (usually 4-7 clauses)	- 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 3 (continued)

General Impression	Focus/Organization	Concrete Language	Elaboration	Mechanics
3 SOME EVIDENCE OF ACHIEVEMENT DEVELOPING WRITER	- controlling idea is not clearly present or easily implied - possible contradiction in position - usually stays on topic, but without discernible plan - possible digressions - may begin or end awkwardly or abruptly	- concrete language may be used with abundant sensory detail, but some details may be inappropriate - details often presented as a list - some writers may use few details or use them inconsistently - details may be adequate in places and absent in other places	- restricted elaboration of one main point (usually 2-4 clauses)	- some minor errors - some major errors - no fewer than 5 combined errors (major and minor) - some errors cause reader confusion
2 LIMITED EVIDENCE OF ACHIEVEMENT EMERGING WRITER	- very limited sense of position on subject - vague organization with little or no planning - may have significant digression - usually no sense of closure on subject - may be brief	- little concrete language - naming may be simply generic	- limited elaboration (at least 1 clause)	- many minor errors - many major errors - many errors cause reader confusion and interference with understanding
1 MINIMAL EVIDENCE OF ACHIEVEMENT INSUFFICIENT WRITER	- no apparent controlling idea - no apparent plan with little unity or coherence - may be too brief to determine organization	- no concrete language	- no elaboration of any point or general statement	- many major and minor errors causing reader confusion - difficult to read

Table 4

Elementary Portfolio Collection Scale

General Competence	Criteria
6 EXCEPTIONAL ACHIEVEMENT EXCEPTIONAL WRITER	- unified, focused compositions - topic or ideas consistently clear; no digressions - typically clear beginnings, middles, ends - transitions typically smooth and logical - details varied, vivid - details consistently support logic or idea - points are often extensively elaborated (8-10 clauses) - mechanical errors are minor and infrequent
5 COMMENDABLE ACHIEVEMENT COMMENDABLE WRITER	- generally well organized according to definite plans - topics or ideas generally clear - typically clear beginnings and ends - most transitions smooth and logical - details generally varied and vivid; metaphors may sometimes be inappropriate - most details consistent with overall plans - in each composition, at least one point is fully elaborated (6-9 clauses) - mechanical errors do not confuse reader, but in each composition there may be several minor errors, one or two major errors
4 ADEQUATE ACHIEVEMENT COMPETENT WRITER	- controlling topics, ideas, or overall plans always present but do not always focus the writing - ending may sometimes be awkward or abrupt - transitions are typically logical but may be repetitive - in most compositions, some moderate elaboration (4-7 clauses) - some use of specific, clear, realistic detail consistent in the overall plans - details are vivid, may on occasion lack depth and/or direct relevance
3 SOME EVIDENCE OF ACHIEVEMENT DEVELOPING WRITER	- topics or overall plans may not be clearly present - possible digressions or elaborations confusing to reader - some transitions - beginnings and endings may be awkward or abrupt - key elements may be unevenly developed or omitted - details used inconsistently - restricted elaboration of one main point (2-4 clauses) - mechanical errors, some minor, some major, which may on occasion confuse reader

Table 4 (continued)

General Competence	Criteria
2 LIMITED EVIDENCE OF ACHIEVEMENT EMERGING WRITER	- topics, ideas or plans may often not be clear - use of supporting details or events may not be logical - may be digressions or overelaboration that significantly interfere with reader understanding - typically little sense of beginnings or endings - few transitions - minimal use of supportive detail; detail may be irrelevant or confusing - many mechanical errors which interfere with understanding
1 MINIMAL EVIDENCE OF ACHIEVEMENT INSUFFICIENT WRITER	- topics may be clear but no overall organizational plans - many digressions or overelaborations or little development altogether - few or no transitions - little sense of beginnings or endings - many mechanical errors interfere with understanding - incomplete sentences

Table 5

Pilot Rubric for Students' Progress

3	Marked, striking increase in competence
2	Moderate increase in competence
1	Some increase in competence
0	No change
-1	Some decline in competence
-2	Marked decline in competence
X	Can't score; portfolio pieces are too varied in genre and content and/or number of pieces is too small

Rating Procedures

Each of our three raters was a teacher experienced in using the 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 did begin the session by scoring and reaching agreement on a training set of 20 samples.

In conducting the narrative scoring, raters were informed that the samples represented a mix of Grades 1, 3, and 4. The material was distributed so that two raters rated each piece independently (in counterbalanced order), and the third rater rated 50% of the samples distributed equally between the other two raters. To prevent rater drift, a check set of 20 papers was included half-way through the scoring session; any disagreements were resolved through discussion that made certain that raters were not changing their criteria for scoring.

For the summaries, one of the raters adapted the narrative rubric and then refined the adaptation with a second rater during an initial training set. These two raters then scored the remainder of this small set independently.

For scoring portfolio collections, the raters requested that we separate and identify the portfolios by grade level. We provided them with an initial set of 6 third-grade portfolios for scheme development. They then scored the remaining portfolios independently, completing the third-grade set, and then proceeding to the fourth and the first grades.

Rater Critique

Recognizing that portfolios have the potential to provide information different from what is intended by standard writing assessments, we asked the raters to critique the appropriateness of the rubric for portfolio scoring and to suggest alternative or supplementary schemes.

Results

Rater Agreement

> *Rater agreement: Can a holistic/analytic scheme be applied to the scoring of classroom samples with the same levels of rater agreement typically reported for standard writing assessments?*

Rater agreement was examined using both percent agreement and correlation coefficients. We computed these just for the material rated independently, excluding ratings assigned after discussion during the training or the check sets. For all remaining analyses, we used scores assigned by raters during both the independent scorings and the training and/or check sets; we treated as equivalent the average of the independent

ratings with the single score achieved through discussion during the training and check sets.

Patterns of rater agreement differed across type of assessment (Classroom Narratives, Classroom Summaries, Standard Writing Assessment, and Portfolio Collections) and across subscales (Tables 6-9). While overall agreement across type of assessment was satisfactory, it was highest for Portfolio Collections and for the Standard Writing Assessment and lowest for Classroom Summaries. Differences in agreement among the subscales were inconsistent across assessment types, and there were no consistent differences among rater pairs in levels of agreement. Pointing to scale stability over time, there was very high agreement between this group of raters and the earlier group (from the summer of 1990) in their ratings of students' Standard Writing Assessments (Table 10).

Table 6

Rater Agreement: Portfolio Collections (*N*=35)

Index	**Agreement**
Pearson correlation coefficients	
Raters 1 and 2	**.76***
Raters 1 and 3	**.74***
Raters 2 and 3	**.94***
Percent agreement ±0	
Raters 1 and 2	**1.00**
Raters 1 and 3	**.97**
Raters 2 and 3	**1.00**
Percent agreement ±1	
Raters 1 and 2	**1.00**
Raters 1 and 3	**1.00**
Raters 2 and 3	**1.00**

*** *p*<.001**

Table 7

Rater Agreement: Classroom Narratives

	Scale			
Index	General Competence	Focus/ Organization	Development	Mechanics
Pearson correlation coefficients				
Raters 1 and 2 (72)	.56*	.44*	.62*	.35*
Raters 1 and 3 (70)	.72*	.64*	.72*	.44*
Raters 2 and 3 (117)	.53*	.37*	.55*	.50*
Percent agreement ±1				
Raters 1 and 2	.95	.93	.97	.84
Raters 1 and 3	.99	.93	.97	.92
Raters 2 and 3	.92	.86	.98	.80
Percent agreement ±0				
Raters 1 and 2	.25	.17	.22	.16
Raters 1 and 3	.28	.15	.23	.21
Raters 2 and 3	.36	.25	.32	.33

* *p*<.001

Table 8

Rater Agreement: Classroom Summaries (*N*=15)

	Scale			
Index	General Competence	Focus/ Organization	Development	Mechanics
Pearson correlation coefficients				
Raters 1 and 3	.41	.36	.49	.09
Percent agreement ±1				
Raters 1 and 3	.89	.84	.95	.95
Percent agreement ±0				
Raters 1 and 3	.47	.16	.32	.37

Table 9

Rater Agreement: Standard Writing Assessment

	Scale			
Index	General Competence	Focus/ Organization	Development	Mechanics
Pearson correlation coefficients				
Raters 1 and 2	.83*	.76*	.74*	.80*
Raters 1 and 3	.84*	.67	.83*	.78*
Raters 2 and 3	.85***	.78***	.78**	.36
Percent agreement ±1				
Raters 1 and 2	1.00	1.00	1.00	1.00
Raters 1 and 3	1.00	.83	1.00	1.00
Raters 2 and 3	1.00	1.00	.92	.80
Percent agreement ±0				
Raters 1 and 2	.63	.38	.25	.38
Raters 1 and 3	.43	.17	.50	.25
Raters 2 and 3	.57	.46	.46	.40

* p<.05.   **p<.01.   ***p<.001.

Table 10

Agreement Between Two Independent Groups of Raters: Standard Writing Assessment (N=18)

	Scale[a]		
Index	General Competence	Focus/ Organization	Development
Pearson correlation coefficients	.72**	.66*	.72**
Percent agreement ±1.0	.97	.94	.97
Percent agreement ±.5[b]	.76	.76	.79

[a] The initial group of raters did not score Mechanics.

[b] Since we were computing the agreement between the *average* ratings of each group, exact agreement was extremely unlikely.

* p<.005.   **p<.001

Validity: Sensitivity of Ratings to Developing Writing Competence

For each type of assessment, do raters' judgments reflect grade level differences in competence?

Table 11 contains descriptive statistics for scores assigned to each type of assessment, by grade level. As shown in the table, we computed means for the Standard Assessment, Classroom Narratives, Classroom Summaries, Classroom Narratives and Summaries combined, and Portfolio Collections. The combined Narrative/Summary score served as an estimate of a portfolio score based on an aggregation of individual sample scores; since the unscored pieces were either poems, which the raters regarded as unscorable, or a letter present in only a few of the portfolios (three of the Grade 3 and one of the Grade 4 portfolios), the Narrative/Summary score represented the bulk of the *scorable* samples presented to raters in each portfolio collection.

For each type of assessment, there were score differences in the expected direction by grade level: For almost all comparisons, the scores of students in the upper grades were higher than those of students in the lower grades. Because there were so few subjects in Grades 1 and 4, and because the number of subjects differed greatly from grade to grade, statistical comparisons are inappropriate. (Exploratory ANOVAs did provide tentative support for most of the grade level differences.)

Table 11

Descriptives

	Scale			
Task	General Competence	Focus/ Organization	Development	Mechanics
		Grade 1		
Classroom Narratives (*N*=5)				
Mean	2.70	2.94	2.37	3.78
SD	.42	.54	.45	.48
Portfolio Collections (*N*=5)				
Mean	3.20			
SD	.84			

Table 11 (continued)

	Scale			
Task	General Competence	Focus/ Organization	Development	Mechanics
		Grade 3		
Standard Writing Assessment (*N*=16)				
Mean	2.59	2.81	2.57	2.90
SD	.79	.83	.89	.82
Classroom Narratives (*N*=23)				
Mean	3.05	2.85	3.19	3.63
SD	.75	.76	.78	.65
Classroom Summaries (*N*=13)				
Mean	2.74	2.81	2.62	3.39
SD	.51	.70	.48	.54
Narratives & Summaries (*N*=13)				
Mean	2.96	2.81	3.05	3.56
SD	.68	.71	.67	.61
Portfolio Collections (*N*=23)				
Mean	3.42			
SD	.85			
		Grade 4		
Standard Writing Assessment (*N*=5)				
Mean	3.64	3.47	3.69	3.36
SD	1.04	1.90	1.01	1.15
Classroom Narratives (*N*=5)				
Mean	3.39	3.20	3.66	3.88
SD	1.40	1.32	1.07	.74
Classroom Summaries (*N*-6)				
Mean	4.36	4.19	4.28	4.06
SD	.80	.76	.87	.70
Narratives & Summaries (*N*=6)				
Mean	3.91	3.72	3.90	4.02
SD	.90	.89	.54	.80
Portfolio Collections (*N*=6)				
Mean	4.00			
SD	.70			

Consistency of Students' Performance Across Writing Contexts; Effects of Assessment Type on Raters' Judgments

> *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?*

Four approaches to data analysis were used to examine relations among students' scores for Standard Writing Assessment, Classroom Narratives, Classroom Summaries, Narratives and Summaries combined, and Portfolio Collections: (a) repeated measures comparisons of students' scores across types of assessments, (b) correlations of scores among types of assessments, (c) cross-classifications of students whose writing was classified as "adequate" versus "inadequate" based on each type of assessment, and (d) cross-tabulations of students' scores across types of assessment. These analyses were restricted to Grade 1 because of the small sample size of Grades 1 and 4.

Repeated measures comparisons. Repeated measures comparisons showed that ratings of student performance differed across type of assessment and task context for each scale except Focus/Organization (Table 12); in addition, Narrative scores were higher than those for Summaries. The results suggest that: (a) students were fairly consistent in their abilities to organize their writing across task contexts that differed markedly in genre, topic, and length; (b) differences in task requirements (students' access to resources, more time, and in some cases the assistance of others) had greater impact on the extent and quality of their compositions' development and mechanics than on focus and organization; and (c) grade level emphasis on narrative writing was associated with Narrative scores higher than Summary scores. (While not reported here because of small sample size, the results for the fourth-grade students also reflected grade level emphases: The fourth graders performed more competently on Summaries than on Narratives, a pattern consistent with their teacher's emphasis on descriptive writing.)

Table 12

Repeated Measure Comparisons for Grade 3

Measure	N	Standard Assessment	Classroom Narratives	Classroom Summaries	Narratives & Summaries	Portfolio Collections
General Competence	(8)	2.90	3.45	2.85		3.58
	(16)	2.59			3.02	3.40*
	(23)				2.96	3.42**
	(23)	2.90	3.45	2.85		
	(8)	2.59			3.02	
	(8)	2.59				3.40*
	(16)					
	(16)					
Focus/Organization	(8)	3.02	3.10	3.06		
	(8)	2.81			2.84	
	(16)					
Development	(8)	2.81	3.66	2.67*		
	(8)	2.57			3.11*	
	(16)					
Mechanics	(8)	3.19	3.84	3.38†		
	(8)	2.89			3.49**	
	(16)					

* p < .05. ** p < .01. †p < .06

Portfolio scores—whether indexed by the Portfolio Collection score or the aggregate Narrative & Summary score—were higher than the Standard Assessment score, indicating that students' classroom work may have been of higher quality than their responses to a standard prompt. If rater bias was a factor, the raters could have perceived the classroom work (which was longer and often illustrated) as more competent than the briefer standard responses. The finding that Portfolio Collection scores were higher than the aggregate Narrative & Summary scores may indicate that the raters assigned the global collection score on the basis of the more competent pieces in each collection.

Correlations of scores across types of assessment for Grade 3. Within each assessment type, most subscale scores were highly intercorrelated, with the exception of those for Mechanics. The infrequency of significant relationships between Mechanics and other scale scores indicates that raters were differentiating the quality of students' compositions from students' skills with grammar, spelling, capitalization, and punctuation (Table 13). Only the intercorrelations for Summaries diverged from this pattern, in that the subscale scores for Development were not correlated with other scores; a possible interpretation is that the brevity of most of the summaries (compared with the narratives) resulted in raters' differing interpretations of the students' task and the expected level of detail.

The Standard Assessment scores were not correlated with scores for most other assessment types. Exceptions were the scattered relationships between some of the subscale scores for the Standard Assessment—itself a narrative—and Classroom Narratives. The absence of a relation between scores for Standard Writing Assessments and Portfolio Collections is particularly interesting in that the portfolios *contained* the Standard Assessment; the result provides additional evidence that the raters were not strongly influenced in their portfolio judgments by the less adequate samples in the collections.

Portfolio Collection scores were correlated only with Classroom Narratives and with Narratives & Summaries combined; Portfolio Collections were not correlated with Classroom Summaries. Since most of the Grade 3 pieces were narratives, the findings are likely to reflect commonality of material as well as a possible tendency on the part of raters to bias their whole portfolio score toward the more competent samples.

Table 13

Correlations Among Measures for Grade 3

		Scale			
Task	*N*	General Competence	Focus/ Organization	Development	Mechanics
		Standard Assessment			
Standard Assessment					
General Competence	(16)		.91***	.85***	.59*
Focus/Organization				.89***	.53*
Development					.32
Mechanics					
Classroom Narratives					
General Competence					
Focus/Organization					
Development					
Mechanics					
Classroom Summaries					
General Competence					
Focus/Organization					
Development					
Mechanics					
Narratives & Summaries					
General Competence					
Focus/Organization					
Development					
Mechanics					
Portfolio Collections					
		Classroom Narratives			
Student Assessment	(16)				
General Competence		.31	.30	.42	.40
Focus/Organization		.35	.37	.47	.50*
Development		.20	.24	.32	.44
Mechanics		.68**	.55*	.70**	.49†
Classroom Narratives	(23)				
General Competence			.92***	.95***	.68***
Focus/Organization				.84***	.71***
Development					.67***
Mechanics					

*p<.05.   **p<.01.   ***p<.001.   †p<.06.

Table 13 (continued)

Task	*N*	Scale			
		General Competence	Focus/ Organization	Development	Mechanics
			Classroom Narratives (cont)		
Classroom Summaries					
General Competence					
Focus/Organization					
Development					
Mechanics					
Narratives & Summaries					
General Competence					
Focus/Organization					
Development					
Mechanics					
Portfolio Collections					
			Classroom Summaries		
Standard Assessment	(8)				
General Competence		.13	.23	-.46	-.17
Focus/Organization		.24	.44	-.28	-.08
Development		.02	.22	-.55	-.25
Mechanics		.53	.48	-.12	.18
Classroom Narratives	(13)				
General Competence		.59*	.56*	.49	.66*
Focus/Organization		.47	.37	.35	.55†
Development		.48	.50	.47	.55†
Mechanics		.06	.09	.19	.53
Classroom Summaries	(13)				
General Competence			.86***	.54†	.39
Focus/Organization				.43	.25
Development					.43
Mechanics					
Narratives & Summaries					
General Competence					
Focus/Organization					
Development					
Mechanics					
Portfolio Collections					

**p*<.05. ***p*<.01. ****p*<.001. †*p*<.06.

Table 13 (continued)

		Scale			
Task	*N*	General Competence	Focus/ Organization	Development	Mechanics
		P & Summaries			
Standard Assessment	(16)				
General Competence		.24	.26	.30	.31
Focus/Organization		.28	.34	.38	.43
Development		.12	.20	.19	.36
Mechanics		.65**	.52*	.64**	.42
Classroom Narratives					
General Competence					
Focus/Organization					
Development					
Mechanics					
Classroom Summaries					
General Competence					
Focus/Organization					
Development					
Mechanics					
Narratives & Summaries	(23)				
General Competence			.92***	.94***	.64***
Focus/Organization				.83***	.66***
Development					.64**
Mechanics					
Portfolio Collections					

**p*<.05. ***p*<.01. ****p*<.001. †*p*<.06.

Table 13 (continued)

Task	*N*	Portfolio Collections
Standard Assessment	(16)	
General Competence		.02
Focus/Organization		.21
Development		.02
Mechanics		.40
Classroom Narratives	(23)	
General Competence		.62**
Focus/Organization		.52*
Development		.67***
Mechanics		.34
Classroom Summaries	(13)	
General Competence		.27
Focus/Organization		.39
Development		.43
Mechanics		.48
Narratives & Summaries	(23)	
General Competence		.60**
Focus/Organization		.54**
Development		.68***
Mechanics		.33
Portfolio Collections		

**p*<.05. ***p*<.01. ****p*<.001. †*p*<.06.

There were some positive relationships between the scores for Classroom Narratives and Classroom Summaries, indicating some commonality of students' writing abilities across these two classroom genres.

Cross-classifications. Using 3.5 as the criterion for adequacy, we classified students' writing as "adequate" or above versus "inadequate" and then compared students' classifications for each type of assessment. Table 14 contains the results.

Table 14

Comparisons of Students' Classifications as “Adequate” or “Inadequate” Writers Based on Each Type of Rating Material

	Classroom Narratives		Classroom Summaries		Narratives/Summaries		Portfolio Collections	
Material & Adequacy	Adequate	Inadequate	Adequate	Inadequate	Adequate	Inadequate	Adequate	Inadequate
Student Assessment								
Adequate	2	0	0	2	2	0	2	0
Inadequate	2	12*	0	6[a]	2	12*	4	10
Classroom Narratives								
Adequate			0	6	6	0	5	1
Inadequate			0	7[a]	0	17	5	12*
Classroom Summaries								
Adequate					0	0	0	0
Inadequate					6	7[a]	6	7[a]
Narratives/ Summaries								
Adequate							5	1
Inadequate							5	12*

Note. Scores for Classroom Narratives, Classroom Summaries, and Narratives/Summaries were the means of scores assigned to all of a student's separate pieces. Adequacy was then defined as 3.5 or greater.

[a] Could not be tested.

* Fisher exact test, one-tailed, $p \leq .05$. ** Fisher exact test, one-tailed, $p < .001$.

Although Grade 3 students' writing was generally rated as inadequate, it was more likely to be rated as adequate when based on Portfolio Collections rather than on either of the Classroom genre sets, the Narratives & Summaries aggregate index, or the Standard Assessment. Thus while all of the children who had written summaries (13) were judged as inadequate writers based on their Summaries, 6 of these 13 children were judged as adequate based on their Portfolio Collections. While 17 of the 23 children who had written narratives were judged as inadequate based on their Narratives, 5 of these 17 were judged as adequate based on their Portfolio Collections. While 14 of the 16 children who completed the Standard Assessment were judged as inadequate, 4 of these 14 were judged as adequate based on their Portfolio Collections.

Children were somewhat more likely to have been rated as adequate based on their Classroom Narratives than on the Standard Assessment (which was also a narrative); of the 14 children rated as inadequate on the Standard Assessment, 2 were rated as adequate on the basis of their Narratives.

Cross-tabulations of students' scores across types of assessment. Cross-tabulations of raters' scores across types of assessment (Table 15) showed raters' judgments of students' competence were often equivalent. If the judgments based on the Standard Assessment differed, they were most often lower than those for Narratives and for Portfolio Collections, but not lower than those for Summaries. Students' summaries were, in fact, composed under conditions somewhat similar to the Standard Assessment—in-class, one-period responses to assigned topics such as "Our Field Trip" or "My Vacation." Judgments based on the Classroom Narratives were most often higher than those for Classroom Summaries. If judgments based on Portfolio Collections differed, they were almost always higher than judgments based on any other type of assessment and often higher than the aggregate Narrative/Summary score; thus, this analysis shows that even students' average scores for their more competent writing genre were often less than the single scores assigned the entire Portfolio Collection.

Table 15

Comparisons of Students' Classifications Based on Each Type of Material

	Student Assessment			Classroom Narratives			Classroom Summaries			Narratives & Summaries		
Material	Higher	Equal	Lower	Higher	Equal	Lower	Higher	Equal	Lower	Higher	Equal	Lower
Classroom Narratives	2	4	10									
Classroom Summaries	3	2	3	10	3	0						
Narratives & Summaries	2	4	10	0	23	0	0	3	10			
Portfolio Collections	0	8	8	2	13	8	0	4	9	2	13	8

Note. Scores used for Classroom Narratives, Classroom Summaries, and Narratives & Summaries were the means of scores assigned to all of a student's separate pieces.

Raters' Critiques

Problems and Prospects for Portfolio Assessment

What are raters' opinions of the utility of a holistic/analytic rubric for portfolio scoring?

We created for scoring purposes a portfolio structure that addressed what we considered to be important first questions regarding the feasibility of portfolio assessment for evaluation of students' writing competence. The portfolios contained final drafts of all assignments, sequenced by date. The raters' focus group discussions about their scoring experience centered on one central issue: *Analytic rubrics have potential, but coordinated portfolio structures need to be designed to provide scorable information.* Below, we summarize those raters' views that bear more directly on the scorability of the portfolios that we presented them. (Some of these issues were also raised in focus groups reported by Meyer, Schuman, & Angello, 1990.) In the final section of our paper, we return to these and additional issues raised by the raters, in order to outline implications of our study for future directions in portfolio assessment.

The Need to Structure the Portfolio Contents

Mix of genre and topic. Portfolios were reflections of the students' classroom assignments, and, as such, there was a considerable mix of genre and topic in each collection. The raters felt that the mix of genres obscured evidence of change over time in writing quality. 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. Although raters did feel able to assign a General Competence score to the mixes that we presented, they did not assign any subscales or a progress score and were unable to use the presented portfolio collections as a basis for further scheme development.

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 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, particularly 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. We return to these issues at greater length in the final discussion section of the paper.

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. Our raters had never evaluated word-processed writing samples, and were not using computers in their own classrooms. Their comments indicated some misperceptions of the functions that computers serve in students' writing. They 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 with 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

The 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. The raters 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. (Indeed, in our prior assessments of ACOT students' writing, we routinely mixed grade levels in order to evaluate differences in performance across grade levels.) Our raters felt strongly that criteria for assigning scores differed for each grade level. 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? The answer to this question depends upon whether one perceives writing as a cross-age developmental process and whether one wishes to couple tightly scoring rubrics to traditional conceptions of grading (e.g., 6=A, 5=B, etc.).

Matching Design and Purpose

Conflicts between concepts of "portfolios" and the designrequirementsfor 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, and 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 a necessary inclusion. 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 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. To illustrate, Table 16 contains the sketch they provided for each of the scale points. 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, they suggested using portfolios to assess additional dimensions of student performance. 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.

Table 16

Shopping List of Suggestions

Commendations	Recommendations
Description/Detail	
• uses clear images	• try using descriptive words
- is vivid	• use colorful words
- is concrete/sensory	• help reader feel, hear, see, smell,taste
• has willingness to risk	• try something you haven’t tried before
• uses detail consistent with intent	• make sure your details match your subject
- helps reader to visualize	• compare things to other things
- shows, doesn’t tell	• help reader feel, hear, see, smell, taste
Organization	
• subject clear	• make sure your reader knows what you’re writing about
• logical flow, beginning, middle, end	• include a beginning, middle and end
• uses varied transitions when appropriate	• choose different ways to move your writing piece along
• orients reader to subject	• give your reader a clear idea of your subject
• provides reader with sense of closure	• tie the parts of your writing piece together at the end
• most sentences directly develop topic	• choose ideas that relate to the topic

Table 16 (continued)

Commendations	Recommendations
Mechanics	
• sentence construction is correct	• make each sentence one complete idea
• spelling developmentally appropriate	• try to be aware of your spelling so others can read your writing
• 3rd grade and up paragraphing developmentally appropriate	• 3rd grade and up use the paragraphing skills you've learned
• punctuation developmentally appropriate	• try to be aware of punctuation so others can understand your writing
• capitalization developmentally appropriate	• following capitalization rules you have learned
• usage developmentally appropriate and does not cause reader confusion	• make sure you're correct so your reader can understand you

Summary and Interpretation

The purpose of our study was to examine the feasibility of evaluating students' writing competence with a holistic/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 more highly 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 a holistic/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 or by the computer?

Discussion and Implications

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 competence 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 can 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 possibilities 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. Methods of portfolio use must be created 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. What is needed are multiple prototypes suited to the diverse needs of schools.

References

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. (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.

Gentile, C. A. (Forthcoming). *The writing students do in school: The 1990 NAEP portfolio study of fourth and eighth graders' school-based writing*. Princeton, NJ: Educational Testing Service.

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.* Keystone Writing Project, Fort Worth Independent School District, Fort Worth, Texas.

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*, pp. 8-11.

Moss, P. A., Beck, J. S., Ebbs, C., Herter, R., Matson, B., Muchmore, J., Steele, D., & Taylor, C. (1991, April). *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, IL.

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.

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

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 & Associates.

Appendix A

Writing Assignment Story

Writing Assignment

Story

A Very Special Memory

Directions: Think of a very special memory, something that happened to you that you will never forget. Write a story about what happened to you.

In your story, be sure to:

- Tell what happened and the order in which it happened.

- Give details about the situation, people, and events. Also tell how you felt about them.

- Organize your story carefully.

Before you begin, write your name, grade, and school name at the top of the page. Write your name on the top of any additional pages you use.