Issues in Portfolio Assessment: The Scorability of Narrative Collections

CSE Technical Report 410

John R. Novak, Joan L. Herman,
and Maryl Gearhart

May 1996

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
Graduate School of Education & Information Studies
University of California, Los Angeles
Los Angeles, CA 90024-1522
(310) 206-1532

Copyright © 1996 The Regents of the University of California

The work reported herein was supported in part by 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.

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

Abstract

This report illustrates techniques for establishing the reliability and validity of assessments of student writing. Raters scored collections of elementary students' narrative writing with the holistic scales of two rubrics—a new rubric designed for classroom use and known to enhance teacher practice, and an established rubric for large-scale writing assessment. Comparisons of score reliabilities were based on three methods: percent agreement, correlations between rater pairs, and generalizability studies. Comparisons of the evidence for validity of scores were based on: (a) correlations of scores with results from two other methods of writing assessment, (b) developmental patterns across grade levels, and (c) consistency of decisions made across methods of assessment. Results were mixed, providing good evidence for the reliability and developmental validity of the new rubric, while correlational patterns were not clear. The discussion addresses the importance of establishing performance-based assessments of writing that are both technically sound and usable by teachers.

ISSUES IN PORTFOLIO ASSESSMENT:
THE SCORABILITY OF
NARRATIVE COLLECTIONS[1]

John R. Novak, Joan L. Herman, and Maryl Gearhart

CRESST/University of California, Los Angeles

Advocates for assessment reform argue that performance-based assessments can provide sound data on educational attainment as well as information to practitioners that motivates improvements in educational practice (Resnick & Resnick, 1989). These dual goals for reform are ambitious and complex, and their realization requires careful research on both the technical qualities of new assessments and their instructional impact. In this paper, we illustrate analytical techniques for establishing the technical qualities of a performance measure of student writing, a measure whose use by teachers has previously been shown to enhance teachers' knowledge and the quality of their instruction (Gearhart & Wolf, 1994; Gearhart, Wolf, Burkey, & Whittaker, 1994; Wolf & Gearhart, 1995).

The measure we have selected for study is a writing assessment—raters' judgments of students' competence with narrative writing based on the application of a new rubric to collections of students' narratives. Further explanation of this measure requires both definition of "collection" and background on the content of the rubric.

"Collection": A Laboratory Model of Portfolio Contents

Collections of writing are considered here as a special case of a class of new performance assessments known as "portfolio assessments." Although models of portfolio assessment differ, it is common practice that students' classroom work and their reflections on that work are assembled as evidence of growth and achievement. The goal is to produce richer and more valid assessments of

[1] The work reported here was collaborative. The authors' names are listed alphabetically in reverse. Maryl Gearhart and Joan Herman had greater responsibility for the design and conduct of the study, John Novak for the design, conduct, and write-up of the data analyses. We thank Shelby Wolf for the *Writing What You Read* rubric and for rater training.

students' competencies than are possible with traditional testing (e.g., Calfee & Perfumo, 1992; Camp, 1993; Freedman, 1993; Hewitt, 1991, 1993; Hiebert & Calfee, 1992; Hill, 1992; LeMahieu, Eresh, & Wallace, 1992; LeMahieu, Gitomer, & Eresh, in press; Mills & Brewer, 1988; Moss, 1992; Murphy & Smith, 1990; O'Neil, 1992; Paulson, Paulson, & Meyer, 1991; Reidy, 1992; Sheingold, 1994; Simmons, 1990; Tierney, Carter, & Desai, 1991; Valencia & Calfee, 1991; Vermont Department of Education, 1990, 1991a, 1991b, 1991c, 1991d; Wolf, Bixby, Glenn, & Gardner, 1991). However, little is known regarding the capacity of portfolio assessments to support judgments that are valid for large-scale purposes.

The limited research that exists on the technical quality of large-scale portfolio assessment is based primarily on studies of pioneering district and state level projects, including Kentucky (e.g., O'Neil, 1992; Saylor & Overton, 1993; Stroble, 1993), Pittsburgh (LeMahieu et al., 1992; LeMahieu et al., in press), and Vermont (e.g., Koretz, McCaffrey, Klein, Bell, & Stecher, 1993; Koretz, Stecher, Klein, & McCaffrey, in press; Koretz, Stecher, Klein, McCaffrey, & Deibert, 1993). (For reviews of this research literature, see Herman & Winters, 1994, and Herman, Gearhart, & Aschbacher, in press). Interpretation of the results of these investigations is made difficult by variations among the project portfolio models, models that differ in their specifications for contents, for rubrics, and for methods for applying the rubrics. An alternative to field-based investigations is to create laboratory portfolio models that control variation in contents and in scoring methods, and this was our choice. Our laboratory study addressed two issues in large-scale portfolio assessment: (a) the assessment of multiple versus single samples of writing, and (b) the technical quality of rubrics.

Writing What You Read: A Rubric Designed to Support Narrative Instruction

Our goal was to demonstrate the technical soundness of raters' judgments of narrative collections based on a new rubric for assessment of narrative writing. The design of the *Writing What You Read* (WWYR) narrative rubric was prompted by the need for assessment tools that can enhance teachers' understandings of narrative and inform instruction (Wolf & Gearhart, 1993a, 1993b, 1994); the contents of existing rubrics for large-scale narrative assessment were not consistent with current language arts frameworks. The WWYR rubric differs from most narrative rubrics in its narrative-specific content and its developmental framework: It contains five analytic subscales for Theme, Character, Setting, Plot, and Communication, and a sixth holistic scale for Narrative Effectiveness that integrates key concepts from the subscales. An earlier technical study provided evidence of the reliability and validity of all six scales when raters used them for scoring single narrative samples (Gearhart, Herman, Novak, Wolf, & Abedi, 1994; Gearhart, Novak, & Herman, 1994).

For this study, raters applied just the holistic Narrative Effectiveness scale to the collections (Table 1) to capture overall strengths and weaknesses.[2] To provide comparative evidence of technical quality, WWYR results were compared with judgments made with the holistic scale of an existing rubric that was designed for large-scale narrative assessment and has consistently demonstrated sound technical capabilities (Table 2) (Baker, Gearhart, & Herman, 1991; Gearhart, Herman, Baker, & Whittaker, 1992; Herman, Gearhart, & Baker, 1993). (Note: Table 2 displays the subscales, because raters utilizing this rubric are asked to reflect on their subscale judgments when making the General Competence judgments).

Study Design: A Comparative Approach

The goals of our study were twofold—to provide evidence of the validity of a new performance measure, and to illustrate methods for providing that evidence. Our design was comparative: To examine the usefulness and technical quality of the WWYR rubric for scoring collections of student writing, WWYR scores for the collections were compared with results from two other measures: an on-demand, direct writing assessment (students' responses to a standard prompt within a time limit), and the mean of raters' judgments of the individual pieces of classroom work. In addition, these WWYR scores were compared with scores for the same measures derived from application of the holistic scale of a second narrative rubric.

[2] We also examined the potential of the original analytic scales to support consensus in (a) raters' qualitative judgments of a student's strength or weakness, using the analytic scales as a framework (selecting one scale as the strength and another scale for the weakness), and (b) raters' commentary on the collection using the language and constructs contained in any of each rubric's scales. These results are reported in Gearhart, Novak, and Herman (1994).

Table 1

WWYR holistic scale for overall narrative effectiveness: How are features integrated?

1. A character suspended without time, place, action, or conflict. More a statement than a narrative.	*There was a little girl who liked rainbows.* *Poor little Cyclops. He had one eye.*
2. Action-driven narrative written in list-like statements. Character(s) and setting minimal. Plot minimal or missing key pieces in sequence, conflict, or resolution.	*Sleeping Beauty has a prince. She had a balloon and a kite. The sun was very beautiful and shining. She went to a party and she had fun. She had a party dress on and her prince.* *Once there was a little girl. And she was 10 years old. And she was very beautiful. A big bear came out of the forest and she ran deep in the forest. Her name is Amelia. But he was going for Amelia. The little girl was very scared. But then she was happy.*
3. One episode narrative (either brief or more extended) which includes the four critical elements of problem, emotional response, action, and outcome. One or more of these elements may be skeletal. The characters and setting are related but often fairly stereotypical, as is the language which describes them.	See The Dragon Fight and The True Three Little Pigs in the Guidebook. A fable would fit here. *One there was a little girl. Her name was Ashley. She was very pretty. She had red hair and freckles. She also had beautiful brown eyes like brown lakes. Anyway...she was a princess that lived in a golden castle. Her father was the king of the land.* *Oh! I forgot! Ashley had a big sister that was not mean. Her name was Lindsey. And she was just as beautiful as Ashley, but she had brown hair.* *Now the real problem was the grandma. She did not like the children. She thought they were spoiled brats. But the children loved their grandmother.* *It so happened that the grandmother had made a plan so the next day the children would die. And this is how it turns out.* *Well, you see, this woman was not the ordinary grandmother. She actually was a witch. Anyway, she decided to have them go and take a walk in the forest. Then she put a pretty flower out in the path. She knew they would notice it. (If you touched the flower and then touched your hair without washing your hair before two day's time you would die!)* *The next day the girls took a walk in the forest and everything was going as the witch had planned except a couple of drops of water landed in the place where the flower had touched the children's hair.* *When the children came home, the grandma was so angry to see them alive that she jumped off a cliff and was never seen again.*

Table 1 (continued)

4. More than one episode narrative with greater insight into character motivation. Beginning revelation of theme on double levels (both implicit and explicit), and setting is more essential to the tale. Language more detailed, more suited to the narrative, and offers careful transitions.	See THE SEVEN CHINESE BROTHERS (from the youngest's point of view) in the Guidebook. Examples from the story appear under Character and Communication. *The True Story of Cinderella—Dedicated to all the badly treated, beautiful maidens of the world. And the beautiful Fairy godmothers that help them.* *Once upon a item, long ago and far away, there lived Cinderella, and her two ugly step-sisters and one step-mother. They lived in Hollywood in the biggest castle ever made and of all people Cinderella was the poor little servant.* *One night Cinderella had more work than usual. She had to sew dresses and put make-up on her two step-sisters and her ugly mean step-mother. They were going to the prince's ball. The prince was to find a wife. When her step-sisters and step-mother left Cinderella, she started to cry. She wanted to go with her step-mother and step-sisters. All of a sudden a big puff of smoke filled the air and here I am.* *I said that I was her fairy god mother. I am going to help her go to the ball and dance with the prince for the whole night. But as Cinderella turned her head I saw how desperate she really was. But I felt that a man just wants someone to do their dishes and their dirty work for them. Still, she was deeply in love.* *This was where the magic comes in. I took the apple from the table and waved my magic wand above my head and the apple turned into a magical carriage. I took my magic wand and waved it over Cinderella's head and said, "Tirn this filthy little maid into a beautiful princess."* *I took the ants off the other fruit and turned them into horses for the ride there. I looked at her. She was the most beautiful woman I ever saw. Then Cinderella asked, "Why didn't you come before?"* *"I was busy babysitting Goldilocks."* *Then Cinderella and I stepped into the carriage, and we rode into the night. On the way there I told her that she would have to be back by midnight, or the magic will wear out, and she would be the same dirty little girl that she was before. When they got there I changed her ugly step-sisters and step-mother into frogs. Cinderella danced with the prince for the rest of the night. The next day they got married. They lived happily ever after.*

Table 1 (continued)

5. Multi-layered narrative with connected episodes. Character and setting description are detailed and sometimes symbolic to reveal intention, motivation, and integration of individuals with time and space. There is evidence of some risk-taking in plot manipulation (e.g. efforts to foreshadow or embed subplots) and experimentation with language (e.g., figurative language, word play).	*Once there was a king and queen who lived in a golden castle of great beauty, but they had no children. Finally, they had a daughter. They had a splendid feast and they invited all the fairies to court except the eldest fairy because she was a wicked witch.* *When it was time to give the wishes, the eldest fairy stormed in and said, "I curse the child!" Her voice sounded like stones falling from a cliff. "She shall be ugly and when she is fifteen she shall look into a mirror and die!"* *After the wicked witch left, the youngest fairy said, "She shall not die, but just faint for 100 years. However, I cannot change the ugliness. My little wand cannot overpower the eldest fairy." So the king broke all the mirrors in the castle.* *As the ugly princess grew up, it was very hard because everybody in the court teased her. Yet, the servants in the castle loved her as they would their own daughter.* *Time went by and the ugly princess turned fifteen and she decided that she would explore the castle. She went into a tower and there she saw an old woman putting clips into her hair while staring into an odd square of glass that reflected the old woman's face.* *The ugly princess said, "May I try?" She took a clip, and when she stepped before the mirror, she saw her horrible face and fell in a faint to the floor. The witch laughed and said, "I've got you now!"* *Soon, however, the little fairy came and picked up the princess and laid her on a little bed where she slept for a hundred years. But the wicked witch's magic was so powerful that everyone in the castle fell asleep too.* *At the end of the hundred years, an unattractive prince was riding by on a disgusting-looking horse, when he chanced to see a torn up flag fluttering from the tip of a distant tower.* *Then he stopped and remembered a story he had heard when he was only a boy about an ugly princess. Since he hadn't had any luck with beautiful princesses during his journey, he decided to try an ugly one.* *He went into the quiet castle. His footsteps echoed in the halls. Nothing stirred. He felt like the walls were holding their breath. Then he saw a tiny stairway and climbed it to the tower room. When he entered the room, he saw the Sleeping Ugly. He bent to kiss her, but then he stopped and said, "Should I be doing this." But then he decided even though she was ugly on the outside, she was probably very beautiful on the inside.* *He kissed her and she woke up. They were married in a beautiful green meadow with daisies all around. They had two ugly children and they lived happily ever after in a castle without mirrors for the rest of their lives.*
6. A rich and multilayered narrative with fully integrated, often multifunctional components, and considerable orchestration in communication to illuminate the components. Growth in characters, purposeful point of view, variety of plot techniques, crafted choice of language.	(No examples available.)

Table 2

Comparison Narrative Rubric

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
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 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

We address three questions related to technical quality:

- Can raters score collections of student work reliably using rubrics originally designed for single writing samples?
- What is the relative technical quality of the two rubrics applied to collections?
- What is the evidence supporting the validity of WWYR for judging students' writing performance?

Method

Site

The narrative samples were collected from an elementary school located in a middle-class suburb in California.

Data Sets

There were three data sets: direct assessments, samples of classroom narratives, and narrative collections.

The *direct assessments* were designed by the same school district that currently utilizes the comparison rubric. These performance-based assessments were administered over two days. On the first day, students discussed literature related to the prompt. On the second day, students were encouraged to brainstorm and cluster initial ideas for their narratives prior to drafting their response. Students in Grades 2 through 5 responded to a "magic" prompt, and students in Grade 6 to a sports prompt. Excerpts from the "magic" prompt follow:

> Everyone thinks about how exciting it would be to have magical powers. These powers might be used to create something, to change something, or to make something disappear. . . . Imagine a situation where you wished you had magical powers. You might have been at home, at school, or some other place. Pretend you suddenly had magical powers. Think about what you did, how you felt, and how other people acted who were around you. What amazing things did you do with your powers? . . . Write a story that tells what you (or the character) did when you found out you possessed magic powers. Help the reader to understand the situation, what happened, where and when

> it happened, and the people involved. Also include how you (or the character) felt and why. . . .

The *classroom narratives* were the narratives students wrote for class assignments. Because the entire set of classroom narratives was larger than we could score within our project budget, we reduced the set at each grade level with a proportional sampling from each narrative genre (e.g., folk tale, fairy tale, personal narrative, myth, etc.). We did not reduce the Grade 3 sample, however: All narrative assignments were scored to permit us to compute for each third-grade student a competence index based on the mean of scores for individual classroom assignments.

Grade 3 *narrative collections* contained all of the narratives written by each student, usually within a range of 3 to 6 narratives. For Grades 2, 4, 5, and 6, narrative collections were constructed to contain 3 to 6 narrative pieces, sequenced by date. (Thus a few collections with only one or two narratives were eliminated, and a few collections with more than six narratives were reduced in size). Because students' folders varied in their inclusion of writing process materials (drafts, revisions), raters were provided only final drafts to ensure comparability.

Rating Procedures

Raters. The five experienced raters who participated had scored narratives in prior studies (Gearhart, Herman, et al., 1994) and thus had considerable training and experience with both the WWYR and the comparison rubric.

Rating. The order of scoring was designed to reduce possible interactions of order of scoring and rubric on raters' judgments. At each scoring session, raters scored narratives or collections in sets labeled primary (Grades 1–2), middle (Grades 3–4), or upper (Grades 5–6) elementary levels. The number of middle papers or middle-level narratives was the greatest, and therefore, within any of the sets listed, raters rated one half of the middle papers (or collections) first, followed by primary, upper, and the remaining middle papers. This order of scoring grade levels was intended as a modest control over the possible interaction of scoring order and grade level on raters' judgments.

Each phase of scoring began with study and discussion of each rubric, the collaborative establishment of benchmark papers or collections distributed along the scale points, and the independent scorings of at least three papers or collections until disagreement among raters on any scale was not greater than 0.5. (Raters located ratings at midpoints in addition to the defined scale points.) Training papers for each assessment type (direct assessment, classroom narratives, and narrative collections) and for each rubric were drawn from all grade levels. However, when raters began the scoring of a given level (primary, middle, or upper), they first scored several preselected papers or collections at that level independently, resolved disagreements through discussion, and placed these benchmark papers in the center of the table for reference. A check set of three to eight papers was included halfway through the scoring session; any disagreements were resolved through discussion.

Raters rated material in bundles labeled with two raters' names; at any given time, each rater made a random choice of a bundle to score. Before moving to the next phase of the study, scores were compared, and a third rater rated any paper whose scores on any scale differed by more than one scale point.

Because the collection scoring proved to be quite challenging to the raters and proceeded more slowly than initially expected, only 52 collections were scored. This small sample size precludes us from making strong inferences from the results that follow.

Results

Reliability of Narrative Collection Scores

We examined three indices commonly used to assess reliability of raters' judgments—percentages of agreement, correlations between raters, and generalizability coefficients; each of these has strengths and weaknesses that we discuss below.

Interrater agreement. One indication of the reliability of the scoring process is the level of agreement between the pairs of raters, both with respect to the actual scores assigned (absolute agreement) and to the degree to which they provide similar rank orderings of the papers (relative agreement). The percentages of papers on which raters agree within some criterion range provide rough indices of the absolute agreement between the raters.[3] For this study, percent agreements between rater pairs were computed based on exact, ±0.5, and ±1.0 criteria. Table 3 summarizes the results for the four raters who participated in the narrative collection ratings.

The statistics in Table 3 indicate both promise and a problem: There were consistently higher levels of agreement obtained for the WWYR rubric relative to the comparison rubric, but the small sample sizes for each rater pair—ranging

Table 3

Percent Agreement to Within Specified Criteria for Rater Pairs Scoring Narrative Collections

			Agreement criterion		
Rubric	Rater pair	*N*	±0	±0.5	±1.0
COMP	1–4	17	.24	.47	.82
COMP	2–4	12	.17	.42	.83
COMP	3–4	9	.11	.22	.56
COMP	1–3	11	.18	.36	.82
COMP	2–3	17	.24	.88	.94
COMP	1–2	12	.00	.17	.42
Mean			.16	.42	.73
WWYR	1–4	10	.00	.80	1.00
WWYR	2–4	11	.27	.73	.91
WWYR	3–4	11	.18	.64	.91
WWYR	1–3	5	.40	.60	1.00
WWYR	2–3	14	.21	.86	1.00
WWYR	1–2	11	.45	.64	.82
Mean			.25	.71	.94

Note. COMP = Comparison rubric, WWYR = *Writing What You Read* rubric.

[3] These indices must be interpreted with caution, however, since simulation studies indicate that for scoring ranges such as those used on these rubrics, relatively high levels of agreement to within ±1 scale point may be expected solely on a chance basis. See Gearhart, Herman, et al. (1994): Agreement indices were computed for each of 100 "shuffles" of the raters' scores on a 6-point scale. The averages of the agreement indices over 100 repetitions of this process were .16, .44, and .67 for the exact, ±0.5, and ±1.0 levels of agreement, respectively.

from 5 to 17—make the agreement indices for each rater pair quite unstable. The mean of the indices across all rater pairs should provide a much more stable estimate of the true level of agreement, but, again, strong inferences are not warranted.

Correlations between rater pairs. The degree to which different raters agree in the rank ordering of scores may be assessed through the use of correlations between rater pairs. While classical reliability coefficients are defined as the correlations between parallel forms of the same test, we use them here to assess the stability of a student's ranking across different (parallel) raters. Table 4 reports the correlations between rater pairs for narrative collections scored by both rubrics.

Once again, we see a more promising pattern of agreement for the WWYR rubric as opposed to the comparison rubric, but the small cell sizes make inference difficult. The mean correlation across raters for the comparison rubric is .45, while the corresponding statistic for the WWYR rubric is .69. Given the variation in sample sizes, means weighted by the cell sizes might be deemed more appropriate, and those figures are .46 and .67, respectively. Thus the reliability of the comparison rubric as measured by this index is well below what is desirable for any purposes, while that of the WWYR rubric approaches .7, a level which might be considered adequate for purposes where test performance does not carry high stakes for individuals or schools (or have potential serious negative consequences, such as preventing a student from graduating).

Table 4

Correlations Between Rater Pairs for Narrative Collection Holistic Scores

	Comparison				WWYR		
Rater	1	2	3	Rater	1	2	3
2	.04			2	.35		
	(12)				(11)		
3	.85**	.81**		3	.90*	.73*	
	(11)	(17)			(5)	(17)	
4	.43	-.02	.61	4	.63	.68*	.82*
	(17)	(12)	(9)		(10)	(11)	(11)

Note. *N*s are indicated below the correlations in parentheses.

* $p < .05$. ** $p < .01$.

Limitations of percent agreement and correlation coefficients. While percent agreement and correlation coefficients are relatively easy to compute, their meaning has some problems. For example, in the absence of 100% exact agreement, the researcher is left to her own devices to come up with appropriate standards of acceptability for agreement indices. Further, while correlation coefficients have the advantage of easy interpretability, they provide only information about the stability of *relative rankings* rather than *absolute scores*. Their utility is limited in situations where absolute scores play a critical role in decision making. For example, if students needed to reach a certain cut score (as opposed to being in the top two-thirds of the class) in order to graduate, then we would want to be very sure that their chances of attaining that score do not depend on which rater scores their paper.

Generalizability of narrative collection scores. Generalizability theory (Brennan, 1983; Crocker & Algina, 1986; Shavelson & Webb, 1991) is designed to address the limitations of percent agreement and correlation coefficients. In generalizability theory (G-theory) as applied to these data, we try to compute how much of a student's score is attributable to actual capability (true score), and how much to error (factors unrelated to capability). More technically, in this case the total score variance is partitioned into the variance between collections (true variance), the variance that is due to raters, and the variance due to the interaction between raters and collections (error variance). This partitioning of variance allows us to compute generalizability coefficients that are appropriate to relative decisions (when only rank ordering is of interest) and absolute decisions (such as comparing scores to a cutpoint). A second advantage of G-theory is that it supports recommendations for improving reliability. In classical test theory, the reliability of a test is a function of the test length, and one can always improve reliability by lengthening the test. The analogous procedure here is to increase the reliability of a score by having the collection scored by multiple raters and then averaging their scores. Under the framework of G-theory, we can compute G-coefficients that tell us how the reliability of scores is likely to be affected by increasing the number of raters.

The results of these analyses are presented in Table 5. The pattern of results here parallels those found in the earlier analyses, in that the WWYR rubric seems to perform somewhat better as a method for scoring narrative collections than

Table 5

Results of the Generalizability Study for the Narrative Collection Scores

				Generalizability coefficients			
	Variance components			1 Rater		2 Raters	
Scale / Rubric	Person	Rater	Error	Relative	Absolute	Relative	Absolute
Comparison	0.4895	0.1640	0.3699	0.57	0.48	0.73	0.65
WWYR	0.3619	0.0483	0.1930	0.65	0.60	0.79	0.75

does the comparison rubric. These results also provide some evidence for the adequate reliability of WWYR scores of narrative collections based on aggregates of two raters' scores, with relative generalizability of such scores in the vicinity of .80. We repeat that these inferences are based on a very small sample size.

Reliability of Direct Assessments and Classroom Narrative Assignments

Findings comparing the reliability and generalizability of the WWYR and comparison ratings of the direct assessments and individual classroom narrative assignments have been previously reported (Gearhart, Herman, et al., 1994), but we summarize those findings here because of their importance to the validity studies reported in the next section. Gearhart, Herman, et al., 1994, found that estimates of the reliability of WWYR and Comparison holistic scores on classroom narrative assignments were comparable and were in the range of .65 to .68, while the generalizability coefficients for scores based on two raters were in the adequate range (.75 to .81). Similar results were obtained for the scores on direct assessments. We concluded that there is evidence that the reliability of both WWYR and comparison scores for direct assessments and classroom narrative assignments are comparable and reasonably adequate, especially if the scores of two raters are averaged.[4]

[4] Note that the reliability estimates for the classroom narrative assignments in Gearhart, Herman, et al. (1994) are based on a single essay; in the present study, we computed a "Classroom Narrative" score for each Grade 3 student by averaging the scores for each narrative. Thus the reliability estimates for classroom narratives reported in the Gearhart, Herman, et al. study may be interpreted as conservative lower bounds on the reliability of the composite scores.

Validity

In this section, we examine patterns of relationships between WWYR results and other indicators of student writing performance, seeking evidence that can be useful in interpreting the meaning of scores based on the WWYR rubric (Messick, 1992). We focus here on the developmental nature of the rubric, as well as relationships among indicators of both convergent and divergent validity. We present analyses of (a) comparisons of students' scores across grade levels (we would expect scores to increase with grade level, since presumably writing achievement increases with additional years of instruction), and (b) comparisons of scores across assessments and rubrics (e.g., we would expect raters to make similar judgments of the same piece(s) of writing using either rubric, since both rubrics are designed to capture writing competence). Unlike the reliability results reported above, all ratings contributed to these results: Scores were computed as either the mean of raters' independent ratings or the resolved score achieved through discussion during the training and check sets.

Grade level comparisons. Table 6 contains descriptive statistics for each rubric-assessment combination. For each rubric and each assessment, there were score differences in the expected direction by grade level.

To look in greater detail at patterns of score change with increasing grade level, ANOVAs were designed to estimate linear, quadratic, and cubic trends for each rubric/assessment combination (summarized in Table 7). (The WWYR direct assessment [WDA] measure was excluded from this analysis, because only two grade levels were scored.) A linear relationship between grade and performance would indicate that as grade increases, the scores increase at a constant rate: We would expect the same difference between Grades 5 and 6, for example, as between Grades 2 and 3. Higher degree trends would indicate departures from this linearity. For example, an initial increase, followed by a leveling off, and finally another increase, would require a cubic trend. As shown in Table 7, the linear trend was significant for all of the variables. For the Classroom Narratives scored with the Comparison rubric (CCLASS) and Direct Assessments scored with the Comparison rubric (CDA) variables, both the cubic trend and a combined linear and cubic trend were also significant: For example, the CCLASS scores make a sizable increase from Grade 2 to Grade 3, level off through Grade 4, and then increase dramatically again at Grade 6.

Table 6

Descriptive Statistics by Grade Level for Comparison (C) and *Writing What You Read* (W) Scores Assigned to Classroom Narratives (CLASS), Narrative Collections (COLLECT), and Direct Narrative Assessments (DA)

	Grade				
Scale	2	3	4	5	6
CCLASS	1.99	2.50	2.56		3.56
	0.43	0.57	0.41		0.46
	16	*23*	*13*		*17*
CCOLLECT	2.94	3.09	3.39	4.25	3.79
	0.92	0.74	0.57	0.46	0.84
	4	*22*	*7*	*12*	*8*
CDA	2.13	2.33	3.13	3.65	3.98
	0.65	0.73	0.67	0.88	0.72
	47	*54*	*45*	*58*	*42*
WCLASS	2.25	2.47	2.53		2.84
	0.34	0.49	0.35		0.59
	16	*23*	*13*		*17*
WCOLLECT	2.56	2.59	3.02	3.45	3.57
	0.38	0.55	0.62	0.48	0.57
	4	*20*	*8*	*11*	*7*
WDA		2.50	3.19		
		0.51	0.54		
		36	*26*		

Although the descriptives and significance tests support the "developmental validity" of the measures examined, the stability of these trends varied in ways that weaken our interpretations. Each of the means is an estimate, each estimate is subject to sampling variability, and that variability was quite variable across the rubric-assessment combinations and across the grade levels, indicating sizable error in the scores actually observed.[5] We see less of this sampling variability in the WCOLLECT variable, however; this information, combined with

[5] Figure 1 contains plots that show the approximate 95% confidence intervals for the means of the rubric-assessment combinations across grade levels. The CDA variable shows the most stable and interpretable trend, due largely to the relatively large sample sizes for that variable across all grade levels. In contrast, the CCOLLECT variable shows great instability; the inordinately wide confidence bands at the second- and fifth-grade levels can be attributed to the extremely small sample sizes (four and eight collections, respectively). However, we seem to see a more stable pattern for the WCOLLECT variable, even though the sample sizes are equally small.

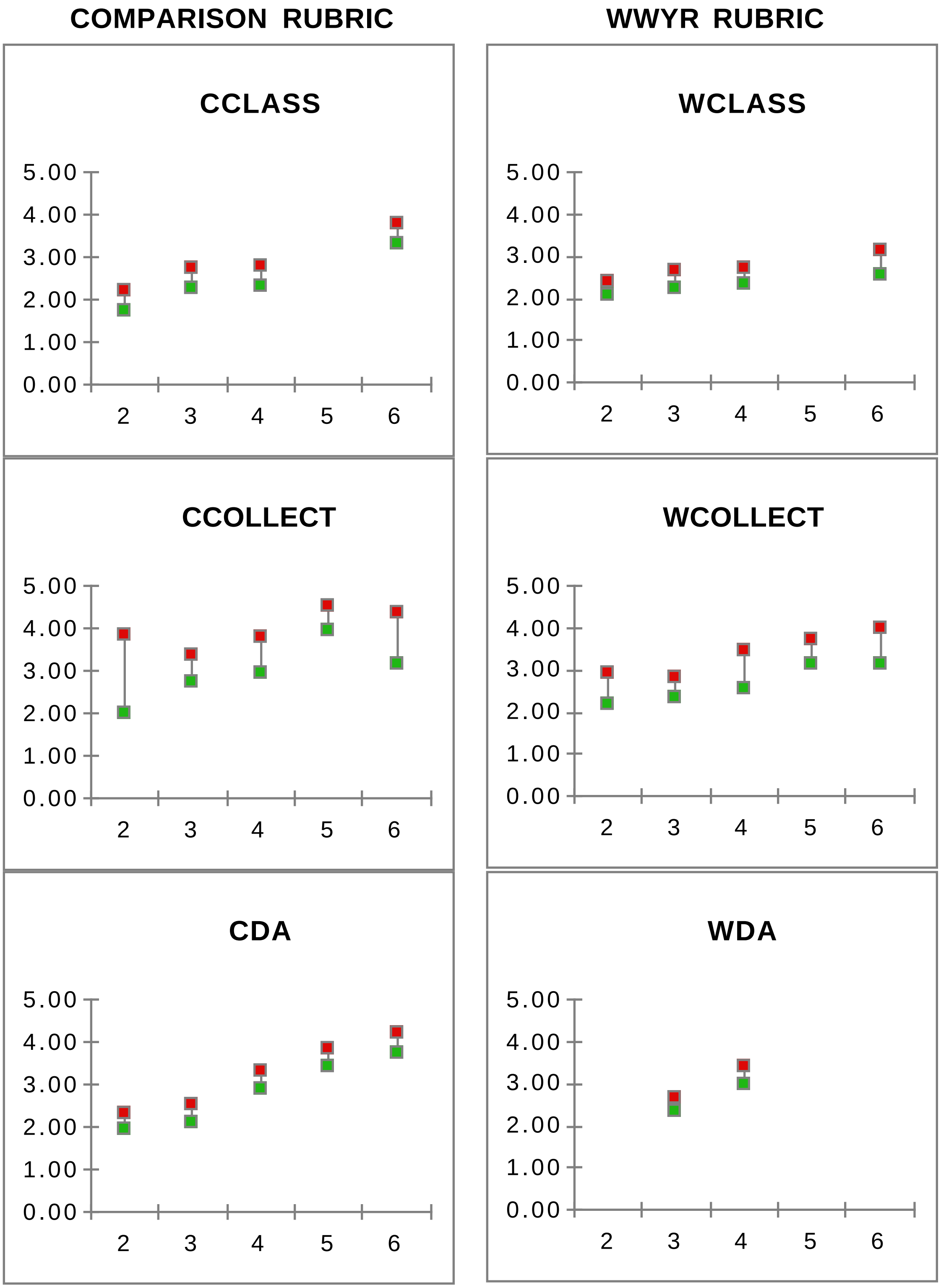

Figure 1. Plots of approximate 95% confidence intervals for the means of the rubric-assessment combinations plotted across grade levels.

Table 7

Summary of ANOVAs for Trend Analyses Across Grade Levels

Scale	*df* Error	Linear	Quadratic	Cubic
Comparison Rubric				
Classroom Narratives	65	88.59	0.01	5.43
CCLASS		0.0001	.9034	.0229
Narrative Collections	48	18.71	2.55	0.06
CCOLLECT		.0001	.1172	.8129
Direct Assessments	238	254.82	0.37	4.53
CDA		.0001	0.5441	.0344
WWYR Rubric				
Classroom Narratives	65	13.67	0.23	0.49
WCLASS		.0004	.6366	.4848
Narrative Collections	45	27.37	0.23	1.13
WCOLLECT		.0001	.6356	.2939
Direct Assessments	60	26.62	NA	NA
WDA		.0001		

Note. Cells contain *F* and *p* values for single degree of freedom contrasts.

the greater reliability of the WWYR rubric over the Comparison rubric as applied to narrative collections, indicates that the WWYR rubric would be more appropriate than the comparison rubric for this purpose.

Comparisons across assessments and rubrics. We examined relationships of scores across assessment types (class assignments, collections, and direct assessments) and rubrics (WWYR, comparison).

Means and variances. Table 8 presents descriptive statistics for students in the third grade; at this grade level, all narratives were scored. We see a remarkable degree of consistency with respect to means and standard deviations across the various rubric-assessment combinations, with the sole exception of the comparison rubric as applied to the narrative collections (CCOLLECT). This finding is consistent with the growing evidence that the comparison rubric is not working well for narrative collections. Statistical tests of differences between scores generally confirm this pattern: Only the CCOLLECT-WCOLLECT comparison showed statistically significant differences, with a repeated measures $F(1, 19) = 10.55$, $p < .01$.

Table 8

Descriptive Statistics for Two Rubrics Applied to Three Assessments, Grade 3 Only

Variable	*N*	Mean	*SD*	Min	Max
CCLASS	23	2.50	0.57	1.35	3.75
CCOLLECT	22	3.09	0.74	1.75	4.50
CDA	52	2.47	0.58	0.75	4.00
WCLASS	23	2.47	0.49	1.42	3.25
WCOLLECT	20	2.59	0.55	1.50	3.50
WDA	36	2.50	0.51	1.50	3.50

Correlations. Table 9 contains all correlations across rubrics and types of material. In such a table we might expect to see the highest correlations when we have (a) the same assessment task scored with different rubrics (the elements on the diagonal of the lower left quadrant of Table 9), and (b) different assessment tasks scored by the same rubric (the elements in the upper left and lower right quadrants). We would also expect to see somewhat lower correlations in instances where different tasks are scored with different rubrics. Unfortunately we do not see any such clear patterns here. The highest correlation that we observe is that between the aggregated scores on the classroom narratives rated by the two rubrics, or WCLASS and CCLASS. But at the same time we also observe that the correlation between WCOLLECT and CCOLLECT is one of the smaller correlations. Similarly, while the correlations among the three scores based on the WWYR rubric are quite consistent and respectable, we see no such consistency among the scores produced with the comparison rubric. In fact, the correlation between CCOLLECT and CDA is the smallest appearing in the table, and indeed is not even significantly different from zero. Since two of the most striking deviations from the expected patterns involve the CCOLLECT variable, we find additional evidence that the comparison rubric is not working for collections.

Consistency of mastery decisions: Background. One potential use of scoring rubrics is to make decisions about students' mastery of skills or competencies. Such a usage requires an initial choice of cutpoint for mastery; thus, for narrative writing, students that score at or above that cutpoint are considered to have mastered the genre, and students scoring below are judged to be nonmasters.

Table 9

Correlations, p-Values, and Sample Sizes Between Scales and Across Rubrics, Grade 3 Only

Measure	CCLASS	CCOLLECT	CDA	WCLASS	WCOLLECT	WDA
CCLASS						
CCOLLECT	.78 .000 *22*					
CDA	.62 .002 *22*	.37 .094 *21*				
WCLASS	.88 .000 *23*	.77 .000 *22*	.71 .000 *22*			
WCOLLECT	.64 .003 *20*	.54 .014 *20*	.78 .000 *19*	.74 .000 *20*		
WDA	.52 .020 *20*	.43 .058 *20*	.72 .000 *35*	.74 .000 *20*	.71 .001 *18*	

When we speak of consistency of decisions we are referring to the degree to which decisions made under different conditions of measurement (i.e., different raters, different occasions, or different rubrics) agree. The results of such paired decision processes can be summarized in a two-by-two contingency table, as exemplified by Table 10.

The cases that fall into the cells on the main diagonal represent those cases in which the decision based on Assessment Method 2 was consistent with decisions based on Assessment Method 1. Those cases in the upper right cell are judged to be masters by Method 2, contrary to their true condition, and hence may be labeled “false positives” for Method 2 relative to Method 1; and similarly the cases in the lower left cell can be labeled “false negatives,” again relative to Method 1. Decisions are consistent to the degree that most of the observations fall into the cells on the main diagonal.

Interpretation of decision consistency analysis requires consideration of two important issues. The first, subject to empirical validation through confirmatory

Table 10

Contingency Table for Examining Decision Consistency

	Method 2	
Method 1	Nonmastery	Mastery
Nonmastery	Consistent	False positive
Mastery	False negative	Consistent

factor analysis, is the underlying assumption that both rubrics are really measuring the same competency, because if they are not, then there would be no reason to expect any kind of decision consistency. However, even if this first assumption is satisfied, decision consistency could be compromised by a second issue, the choice of cutpoints.

Consider a case in which two methods are perfect indicators of some competency and the distributions of the observed scores derived from the two methods are identical except with respect to location. So, for example, the scores on Method 2 might be always one unit larger than those on Method 1. In such an instance there would be a perfect correlation between the two scales, yet, if we were to use the same cutpoint for both scales, these would not provide consistent results. However, if we were to set different cutpoints for the two scales so that the cutpoint for Method 2 is one unit higher than that for Method 1, then we would obtain perfect decision consistency. Standardizing scale scores helps us to deal with this problem.[6]

Measuring decision consistency with the kappa coefficient: Background. The kappa coefficient (Cohen, 1960) is a commonly used statistic for measuring decision consistency. This coefficient may be interpreted as the proportion of decisions that are consistent beyond the proportion that is expected by chance. It may be somewhat loosely compared to a correlation coefficient in that it ranges between -1 and 1. A kappa of 1 may represent perfect consistency, although in our analyses we will see that such a value may be obtained under circumstances

[6] A more complex case would result from a situation in which both methods agree with respect to location (that is they have the same mean) but differ in their variance. In this case, perfect decision consistency can only be achieved if the cutpoint for both methods is set at their common mean; as the cutpoint moves further away from this mean, decision consistency necessarily drops off, perhaps drastically. In the real world we are likely to see cases where the observed scores differ with respect to both location and scale, further complicating the process of setting appropriate cutpoints.

which are less than stable, and so must be interpreted with caution. Negative values of kappa represent levels of agreement below what would be expected by chance.

Decision consistency across the narrative scales and rubrics. We examined the degree to which consistent decisions were made via the various assessment/rubric combinations, and the effect of cutpoints on both the consistency of decisions and the stability of the kappa coefficient. The comparison rubric has a long history of use as a measure of competency for narrative writing, and in prior applications a cutpoint of 3.5 has been used (e.g., Gearhart, Herman, et al., 1994). In the absence of experience using the rubric for narrative collections, we started with that cutpoint. The WWYR rubric has little history as a measure of narrative competency, and so, given the lack of experience and the observed similarities in the distributions of the various rubric-assessment combinations, the same cutpoint was initially used for that rubric as well. Table 11 summarizes the results of the decision consistency analyses. The cells below the diagonal in this table contain the contingency tables for the pairs of rubrics, while the cells above the diagonal contain the kappa coefficients for those pairs.

Data in Table 11 show that the cutpoint of 3.5 is very problematic due to the rarity of decisions of mastery based on this cutpoint. Out of 15 possible pairs of scales with a mean of about 20 observations per pair, there were only 6 observations in which decisions of mastery were made using both scales. If we examine some of the particular cells we can see some of the possible pitfalls of the decision consistency approach through the kappa coefficient in situations where mastery decisions are very rare. First, look at the cells involving the WWYR mean scores for classroom narrative assignments (WCLASS) (the first column of contingency tables and the first row of kappa coefficients). Note that all of the kappa coefficients for these pairs are zero. This will always be the case in which exactly one of the marginal totals (that is, row or column sums) is zero.

An interesting contrast is provided by the results for the WCLASS-WCOLLECT pair and the CCLASS-WCOLLECT pair. For the former, the kappa coefficient is zero, while for the latter the kappa is 1, yet the only difference in the contingency table is the shift of a single observation from the misclassification category to the joint mastery classification.

Table 11

Cross-Classifications (Below Diagonal) and Kappa Coefficients (Above Diagonal) for Mastery Decisions Based on Cutpoints of 3.5 for the Comparison Rubric and 3.5 for the WWYR Rubric

		WCLASS		*WCOLLECT*		*WDA*		*CCLASS*		*CCOLLECT*		*CDA*	
		0	*1*	*0*	*1*	*0*	*1*	*0*	*1*	*0*	*1*	*0*	*1*
WCLASS	0			.00		.00		.00		.00		.00	
	1												
WCOLLECT	0	19	0			-.08		1.00		.18		-.09	
	1	1	0										
WDA	0	18	0	15	1			-.07		-.18		.44	
	1	2	0	2	0								
CCLASS	0	22	0	19	0	17	2			.15		-.08	
	1	1	0	0	1	1	0						
CCOLLECT	0	14	0	13	0	11	2	14	0			-.32	
	1	8	0	6	1	7	0	7	1				
CDA	0	17	0	14	1	29	2	16	1	10	7		
	1	5	0	4	0	2	2	5	0	4	0		

Adjustment of the cutpoints. This undesirable lack of stability in the alpha coefficient is primarily a function of our chosen cutpoints, cutpoints that set a standard for performance that was not attained by most of the sample. Given the developmental nature of the rubric contents and the early developmental stature of the subjects (Grade 3), the cutpoints were adjusted downward. Table 12 contains results for a decision consistency analysis based on cutpoints of 3.0 for both rubrics. The results for these cutpoints are much more promising, and we see that the majority of the kappa coefficients are within the ranges that we were led to expect based on a special simulation study run especially for this study[7]; that is, we are getting decision consistencies that are consistent with what we would expect based on adequately reliable measures with appropriately comparable cutpoints.

[7] In order to gain some insight into what might constitute a reasonable kappa coefficient, a small-scale simulation was run. Simulated observed score distributions were generated based on an underlying continuous ability distribution so as to emulate the conditions that were found in this study. The observed scores were generated so that they would have a reliability close to .80 and similar means and variances. The cutpoint was set in the upper tail of the distribution of observed scores. Kappa coefficients were generated for each simulated data set. The mean of the kappa coefficients over 100 iterations was .51, and the empirical distribution had a standard deviation of .29 and was noticeably skewed in the negative direction. Figure 2 is a histogram of this distribution. Thus, even if the strong assumptions that we have posited regarding unidimensionality and equivalence of distributions are satisfied, then the expected value of our kappa coefficients would be about .51, and observed values could be expected to fluctuate considerably about that value.

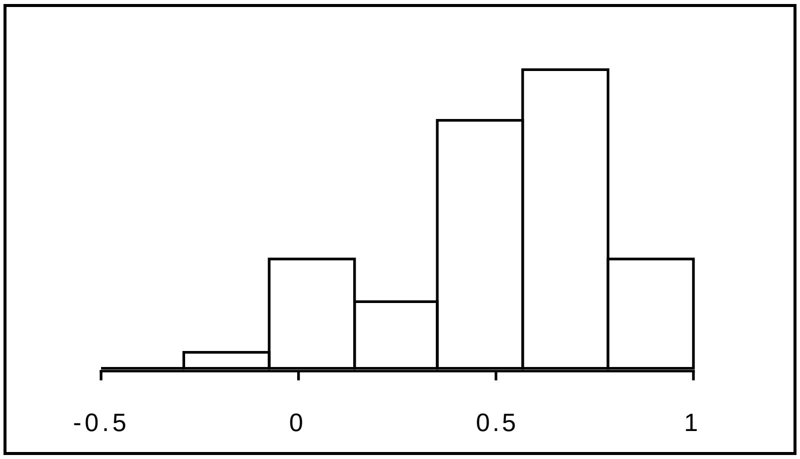

Figure 2. Histogram of the distribution of kappa coefficients from 100 iterations of a simulation process.

Table 12

Cross-Classifications (Below Diagonal) and Kappa Coefficients (Above Diagonal) for Mastery Decisions Based on Cutpoint of 3.0 for the Comparison Rubric and 3.0 for the WWYR Rubric

		WCLASS		*WCOLLECT*		*WDA*		*CCLASS*		*CCOLLECT*		*CDA*	
		0	*1*	*0*	*1*	*0*	*1*	*0*	*1*	*0*	*1*	*0*	*1*
WCLASS	0			.49		.34		1.00		.17		.56	
	1												
WCOLLECT	0	13	4			.61		.49		.34		.77	
	1	0	3										
WDA	0	13	0	11	1			.34		.34		.68	
	1	5	2	2	4								
CCLASS	0	19	0	13	4	13	5			.17		.56	
	1	0	4	0	3	0	2						
CCOLLECT	0	8	0	7	1	7	1	8	0			.29	
	1	11	3	6	6	6	6	11	3				
CDA	0	14	0	11	1	25	3	14	0	7	7		
	1	4	4	1	6	1	6	4	4	1	6		

The major exceptions to the findings are found in those rubric-assessment pairs involving the comparison scores of the narrative collections (CCOLLECT). This is not surprising given the finding that the mean for CCOLLECT was significantly higher than the means for the other measures. Indeed, examination of the contingency tables involving this measure reveals that almost all of the misclassifications are situations in which students were judged as masters based on the CCOLLECT score and were judged as nonmasters using the other score. This is another clear indication that the comparison rubric as applied to the narrative collections functions somewhat differently from the other rubric-assessment combinations. Adjusting the cutpoint upward for CCOLLECT did not alter the results: While the kappa coefficients for CCOLLECT with the other measures were .17, .34, .34, .17, and .29 using the 3.0 cutpoint, a cutpoint of 3.5 for the CCOLLECT scale resulted in respective kappa's of .21, .12, .12, .15, and -.32.

In sum, there is evidence that, if appropriate cutpoints are set, then reasonably consistent decisions can be made regarding the mastery/nonmastery of the narrative writing competency of third-grade students using any of the rubric-assessment combinations, with the sole exception of the comparison rubric scores for the narrative collections (CCOLLECT).

Discussion

The goals of our study were to provide evidence of the validity of a new performance-based writing assessment, and to illustrate methods for providing that evidence. Our questions regarding the technical quality of the new assessment clustered in two categories. One set of questions addressed the meaningfulness of raters' judgments of collections of writing versus single pieces and, as such, represented a strategy for laboratory examination of one of the thorny issues of large-scale portfolio assessment. A second set of questions concerned the roles of rubrics in performance-based assessment. We examined patterns of judgments made with two contrasting rubrics, one designed to support instructional improvement, the other designed for efficient and technically sound large-scale assessment.

We examined reliability of the narrative collection scores using three methods: percent agreement to a range of specified criteria, correlations between rater pairs, and generalizability theory. We discussed the advantages and disadvantages of each approach, concluding that generalizability theory provided the most usable results. Across all analyses, there was a pattern of greater support for the reliability of the WWYR rubric, although the small sample size precluded strong inferences about issues of relative reliability.

There was mixed evidence of validity for the narrative collection scores. Looking for patterns of relationships that supported the meaningfulness of WWYR results, we examined: (a) comparisons of students' scores across grade levels, and (b) comparisons of scores across assessments and rubrics. Support for both rubrics was provided by findings that narrative collection scores increased with grade level, and additional support for the WWYR rubric was provided by the finding that WWYR narrative collection scores for Grade 3 students were consistent with the other two WWYR measures (the direct assessment, and the mean of students' individually-scored classroom narratives). However, the meaningfulness of these WWYR results were challenged by unexpectedly strong relationships between the WWYR narrative collection scores and each of the Comparison measures. Analyses of the consistency of decisions across all rubric-assessment combinations indicated that if appropriate cutpoints are set, then reasonably consistent decisions can be made regarding the mastery/nonmastery of narrative writing competency using the WWYR collection scores, but not the comparison collection scores. Overall, the quantitative results favored the WWYR rubric.

While the small sample size has limited our inferences, the patterns of findings have implications for both of our study questions. First, we produced evidence that multiple samples of writing may be "assessable" provided that certain conditions are in place: These conditions may be the rubric applied to the collections (in our case, WWYR performed better than the comparison rubric), the number of raters who judge the collections (we found from generalizability studies that we needed two), and the judicious use of the scores in ways that are appropriate to evidence regarding the reliability and validity of those scores. Second, we found that different rubrics appeared to frame raters' judgments in different ways. While indices of the technical quality of the WWYR rubric were reasonably comparable across types of material (direct assessment, classroom narratives, collections), the comparison rubric performed quite differently across materials; in addition, raters' judgments of the collections differed for the two rubrics. Thus our findings suggest that choice of rubric can have substantial effect on both the technical quality and the results of a performance assessment. Further research is needed to reveal the ways that raters interpret and apply rubrics to different types of assessment material.

In sum, illustrating a multimethod approach to the technical study of new performance assessments, our study has produced preliminary evidence that, under certain conditions, the holistic scale of the *Writing What You Read* narrative rubric—a rubric previously shown to enhance teachers' understandings of narrative and their methods of instruction—can be used reliably and meaningfully in large-scale assessment of narrative collections. These findings hold promise for the pursuit of large-scale writing assessments that can guide the work of teachers in the classroom and produce technically defensible results.

References

Baker, E. L., Gearhart, M., & Herman, J. L. (1991). *The Apple Classrooms of Tomorrow^{SM}: 1990 evaluation study* (Report to Apple Computer, Inc.). Los Angeles: University of California, Center for the Study of Evaluation.

Brennan, R. L. (1983). *Elements of generalizability theory*. Iowa City, Iowa: ACT Publications.

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

Camp, R. (1993). The place of portfolios in our changing views of writing assessment. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment* (pp. 183-212). Hillsdale, NJ: Erlbaum.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37-45.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart and Winston, Inc.

Freedman, S. (1993). Linking large-scale testing and classroom portfolio assessments of student writing. *Educational Assessment*, *1*(1), 27-52.

Gearhart, M., & Herman, J. L. (1995, Winter). Portfolio assessment: *Whose work is it?* Issues in the use of classroom assignments for accountability. *Evaluation Comment*, 1-16.

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 Research on Evaluation, Standards, and Student Testing.

Gearhart, M., Herman, J. A., Novak, J. R., Wolf, S. A., & Abedi, J. (1994). *Toward the instructional utility of large-scale writing assessment: Validation of a new narrative rubric* (CSE Tech. Rep. No. 389). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

Gearhart, M., Novak, J. R., & Herman, J. L. (1994, November). *Issues in portfolio assessment. The scorability of narrative collections* (Report to OERI, Contract No. R117G10027). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

Gearhart, M., & Wolf, S. A. (1994). Engaging teachers in assessment of their students' writing: The role of subject matter knowledge. *Assessing Writing*, *1*, 67-90.

Gearhart, M., Wolf, S. A., Burkey, B., & Whittaker A. K. (1994). *Engaging teachers in assessment of their students' narrative writing: Impact on teachers' knowledge and practice* (CSE Tech. Rep. 377). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

Herman, J. L., Gearhart, M., & Aschbacher, P. R. (in press). Portfolios for classroom assessment: Design and implementation issues. In R. C. Calfee (Ed.), *Portfolio assessment*.

Herman, J. L., Gearhart, M., & Baker, E. L. (1993). Assessing writing portfolios: Issues in the validity and meaning of scores. *Educational Assessment*, *1*(3), 201-224.

Herman, J. L., & Winters, L. (1994). Portfolio research: A slim collection. *Educational Leadership*, *52*(2), 48-55.

Hewitt, G. (1991). *Leading and learning: A portfolio of change in Vermont schools* (Technical Report). Vermont: Governor's Institutes.

Hewitt, G. (1993, May-June). Vermont's portfolio-based writing assessment program: A brief history. *Teachers and Writers*, *24*(5), 1-6.

Hiebert, F., & Calfee, R. C. (1992). Assessment of literacy: From standardized tests to performances and portfolios. In A. E. Fastrup & S. J. Samuels (Eds.), *What research says about reading instruction* (pp. 70-100). Newark, DE: International Reading Association.

Hill, R. (1992). *Assessments developed in support of educational reform in contrast to assessments developed in support of educational refinement*. Paper presented at the Assessment Conference of the Education Commission of the States, Boulder, CO, June 4, 1992.

Koretz, D., McCaffrey, D., Klein, S., Bell, R., & Stecher, B. (1993). *The reliability of scores from the 1992 Vermont Portfolio Assessment Program* (CSE Tech. Rep. 355). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

Koretz, D., Stecher, B., Klein, S., & McCaffrey, D. (in press). The Vermont Portfolio Assessment Program: Findings and implications. *Educational Measurement: Issues and Practice*, *13*(3).

Koretz, D., Stecher, B., Klein, S., McCaffrey, D., & Deibert, E. (1993). *Can portfolios assess student performance and influence instruction? The 1991-92 Vermont experience* (CSE Tech. Rep. No. 371). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

LeMahieu, P. G., Eresh, J. T. & Wallace, R. C. (1992, December). Using student portfolios for a public accounting. *School Administrator*, *49*(11), 8-15.

LeMahieu, P. G., Gitomer, D. H., & Eresh, J. T. (in press). Portfolios in large-scale assessment: Difficult but not impossible. *Educational Measurement: Issues and Practice*.

Messick, S. (1992). Validity of test interpretation and use. In M. Alkin (Ed.), *Encyclopedia of Educational Research* (6th ed., pp. 1487-1495). New York: Macmillan.

Mills, R. P., & Brewer, W. R. (1988). Working together to show results: An approach to school accountability in Vermont. Montpelier: Vermont Department of Education, October 18/November 10.

Moss, P. A. (1992). Portfolios, accountability, and an interpretive approach to validity. *Educational Measurement: Issues and Practice*, *11*(3), 12-21.

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

O'Neil, J. (1992). Putting performance assessment to the test. *Educational Leadership*, *49*(8), 14-19.

Paulson, F. L., Paulson, P. R., & Meyer, C. A. (1991, February). What makes a portfolio a portfolio? *Educational Leadership*, *48*(5), 60-63.

Reidy, E. (1992, June). What does a realistic state assessment program look like in the near term? Paper presented at the annual ECS/CDE meeting, Boulder, Colorado.

Resnick, L. B., & Resnick, D. P. (1989). Assessing the thinking curriculum: New tools for educational reform. In B. R. Gifford & M. C. O'Conner (Eds.), *Future assessments: Changing views of aptitude, achievement , and instruction* (pp. 37-75). Boston, MA: Kluwer Publishers.

Saylor, K., & Overton, J. (1993, March). Kentucky writing and math portfolios. Paper presented at the National Conference on Creating the Quality School.

Shavelson, R. J., & Webb, N. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage.

Sheingold, K. (1994, September). Designing a large-scale portfolio assessment system for classroom consequences. Presentation at the 1994 CRESST Conference: Getting Assessment Right!, Los Angeles.

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

Stroble, Elizabeth J. (1993). Kentucky student portfolios: Expectations of success. *Equity and Education, 26*(3), pp. 54-59.

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

Valencia, S. W. , & Calfee, R. C. (1991). The development and use of literacy portfolios for students, classes, and teachers. *Applied Educational Measurement, 4,* 333-345.

Vermont Department of Education. (1990, September). *Vermont Writing Assessment: The pilot year.* Montpelier, VT: Author.

Vermont Department of Education. (1991a). *"This is my best": Vermont's writing Assessment Program, pilot year 1990-1991*. Montpelier, VT: Author.

Vermont Department of Education. (1991b). *Looking beyond "The Answer": The report of Vermont's Mathematics Portfolio Assessment Program.* Montpelier, VT: Author [undated].

Vermont Department of Education. (1991c). *Vermont Mathematics Portfolio Project: Resource book*. Montpelier, VT: Author.

Vermont Department of Education. (1991d). *Vermont Mathematics Portfolio Project: Teacher's guide*. Montpelier, VT: Author.

Wolf, D. P., Bixby, J., Glenn, J. & Gardner, H. (1991). To use their minds well: Investigating new forms of student assessment. In G. Grant (Ed.), *Review of Research in Education* (Vol. 17, pp. 31-74). Washington, DC: American Educational Research Association.

Wolf, S. A., & Gearhart, M. (1993a). *Writing What You Read: Assessment as a learning event* (CSE Tech. Rep. 358). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

Wolf, S. A. , & Gearhart, M. (1993b). *Writing What You Read: A guidebook for the assessment of children's narratives* (CSE Resource Paper No. 10). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

Wolf, S. A., & Gearhart, M. (1994). *Writing What You Read:* A framework for narrative assessment. *Language Arts*, *71*(6), 425-445.

Wolf, S. A., & Gearhart, M. (1995, April). Engaging teachers in assessment of their students' narrative writing: Patterns of impact and implications for professional development. Presentation for the panel symposium *The impact of alternative assessments on teachers' knowledge and practice*, presented at the annual meeting of the American Educational Research Association, San Francisco.