DOCUMENT RESUME

ED 462 388 TM 023 887

AUTHOR Gearhart, Maryl; Novak, John R.; Herman, Joan L.
TITLE Issues in Portfolio Assessment: The Scorability of Narrative Collections. Project 3.1: Studies in Improving Classroom and Local Assessments.
INSTITUTION Center for Research on Evaluation, Standards, and Student Testing, Los Angeles, CA.
SPONS AGENCY Office of Educational Research and Improvement (ED), Washington, DC.
PUB DATE 1994-11-00
NOTE 66p.
CONTRACT R117G10027
PUB TYPE Numerical/Quantitative Data (110) -- Reports - Research (143)
EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Educational Assessment; Elementary Education; Elementary School Students; Essay Tests; *Evaluation Methods; *Interrater Reliability; Judges; *Portfolio Assessment; Portfolios (Background Materials); Scoring; State Programs; Test Reliability; Test Use; Test Validity; Testing Programs; Writing (Composition); *Writing Evaluation
IDENTIFIERS Large Scale Programs; *Narrative Text; Writing What You Read

ABSTRACT

Technical questions regarding the reliability and validity of large-scale portfolio assessment were studied which focused on: (1) whether raters can score collections of writing reliably with rubrics designed for single samples; (2) whether ratings derived from different frameworks differ in their capacities to support technically sound assessments of narrative collections; and (3) whether ratings of distinctive narrative assessments characterize groups similarly. The study used 5 raters' judgments of 52 collections of elementary school student writing and was primarily designed to illustrate analytic techniques for addressing each of these questions. Another objective was to evaluate the "Writing What You Read" narrative rubric. The study produced preliminary evidence that the holistic scale of "Writing What You Read" can be used reliably and meaningfully in large-scale assessment of narrative collections. Results support the importance of rubrics designed to capture the qualities of distinctive writing genres. Seven figures and 29 tables present study findings, and 2 appendixes present writing prompts. (Contains 12 references.) (SLD)

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

Final Deliverable – November 1994

Project 3.1 Studies in Improving
Classroom and Local Assessments

Issues in Portfolio Assessment: The Scorability of Narrative Collections

Maryl Gearhart, Project Director

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

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

The work reported herein was supported in part under 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 the Office of Educational Research and Improvement or the U.S. Department of Education.

CONTENTS

ACKNOWLEDGMENTS

Our four raters, Kathy Beneiof, Virginia Espinosa, Rosa Valdes, and Kim Uebelhardt made important and substantial contributions, and John Schacter contributed to the design and coding of raters' commentary on the narrative collections. Our thanks to the teachers, students, and parents who permitted us access to the students' narratives.

ISSUES IN PORTFOLIO ASSESSMENT: THE SCORABILITY OF NARRATIVE COLLECTIONS

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

This report raises technical questions that need to be addressed to assure that portfolios will yield reliable and valid scores for large-scale assessment purposes. First, can raters score collections of narrative writing reliably with rubrics designed for scoring single samples? Second, how do rubrics derived from different frameworks differ in their capacities to support technically sound assessments of narrative collections? Third, do ratings of student writing performance across distinctive assessment contexts (e.g., direct writing assessment vs. collection of classroom writing) categorize groups similarly? Designed to illustrate analytic techniques for addressing each of these questions, this paper draws on findings from a small study of raters' judgments of students' narrative writing. The work reported here builds on two previous lines of investigation—technical studies of portfolio assessment (Gearhart, Herman, Baker, & Whittaker, 1992; Gearhart, Herman, Baker, & Whittaker, 1993; Gearhart & Herman, in press; Herman, Gearhart, & Baker, 1993; Herman & Winters, 1994), and technical studies of a new narrative rubric designed to impact instruction (Gearhart, Herman, Novak, Wolf, & Abedi, 1994; Gearhart, Wolf, Burkey, & Whittaker, 1994; Gearhart & Wolf, 1994; Wolf & Gearhart, 1993a, 1993b, 1994). We begin by reviewing prior findings.

PORTFOLIO ASSESSMENT

In an initial technical study of portfolio assessment (Gearhart et al., 1992; Herman et al., 1993), we examined the feasibility and meaningfulness of evaluating students' writing competence with ratings of (a) a direct assessment, (b) separately scored samples of classroom writing, and (c) portfolio collections. The rubric contained both a holistic scale and analytic scales for differentiated components of writing competence (Table 1); however,

Table 1

Conejo Valley 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 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 1 (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

raters were able to assign only the holistic scale to the portfolios. Results demonstrated that, for all three types of material, raters achieved adequately high levels of agreement, and ratings discriminated meaningfully among grade level and genre differences in students' competence. The generally satisfactory levels of agreement on the holistic scale for portfolio assessment were particularly noteworthy in the context of our raters' perceptions of the difficulty of rating a mix of genres and assignments without knowledge of the assignment expectations or instructional support. The finding that portfolio scores based on holistic judgments of the entire collections were somewhat higher than scores based on aggregates of individually scored samples (the same samples that were contained in the portfolios) raised an issue about the meaning of portfolio scores derived from differing statistical procedures.

The current work largely replicates the design of this early study of portfolio assessability, but extends the work in two ways. First, we restricted the current datasets to collections of narratives only, hoping to reduce some of the complexities raters face when rating portfolios containing multiple writing genres. Second, we examined an additional variable—the contribution of rubrics to rater agreement and to the meaningfulness of the narrative collection scores. We asked raters to apply two different narrative rubrics to the collections—the same rubric used in the initial study (Table 1), and a new rubric derived from current literacy sources that has shown some capability to support large-scale assessment of narrative samples (*Writing What You Read,* Figure 1) (Gearhart, Herman, et al., 1994). Third, we asked the raters to score the collections with an exploratory rubric for narrative progress (Tables 2 and 3).

Table 2

Writing What You Read Progress Rubric

-1	Some decline in narrative effectiveness
0	No change in narrative effectiveness
1	Slight increase in narrative effectiveness
2	Moderate increase in narrative effectiveness
3	Marked, striking increase in narrative effectiveness
N/A	Can't score

Narrative Rubric

Theme

explicit ⟷ implicit
didactic ⟷ revealing

- Not present or not developed through other narrative elements
- Meaning centered in a series of list-like statements ("I like my mom. And I like my dad. And I like my....") or in the coherence of the action itself ("He blew up the plane. Pow!")
- Beginning statement of theme—often explicit and didactic ("The mean witch chased the children and she shouldn't have done that."); occasionally the theme, though well stated, does not fit the story
- Beginning revelation of theme on both explicit and implicit levels through the more subtle things characters say and do ("He put his arm around the dog and held him close. 'You're my best pal,' he whispered.")
- Beginning use of secondary themes, often tied to overarching theme, but sometimes tangential; main theme increasingly revealed through discovery rather than delivery, though explicit thematic statements still predominate
- Overarching theme multi-layered and complex; secondary themes integrally related to primary theme or themes; both explicit and implicit revelations of theme work in harmony ("You can't do that to my sister!", Lou cried, moving to shield Tasha with her body.)

Character

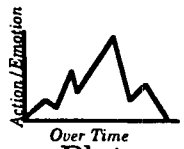
flat ⟷ round
static ⟷ dynamic

- One or two flat, static characters with little relationship between characters; either objective (action speaks for itself) or first person (author as "I") point of view
- Some rounding, usually in physical description; relationship between characters is action-driven; objective point of view is common
- Continued rounding in physical description, particularly stereotypical features ("wart on the end of her nose"); beginning rounding in feeling, often through straightforward vocabulary ("She was sad, glad, mad.")
- Beginning insights into the motivation and intention that drives the feeling and the action of main characters often through limited omniscient point of view; beginning dynamic features (of change and growth)
- Further rounding (in feeling and motivation); dynamic features appear in the central characters and in the relationships between characters; move to omniscient point of view (getting into the minds of characters)
- Round, dynamic major characters through rich description of affect, intention, and motivation; growth occurs as a result of complex interactions between characters; most characters contribute to the development of the narrative; purposeful choice of point of view

Setting

backdrop ⟷ essential
simple ⟷ multi-functional

- Backdrop setting with little or no indication of time and place ("There was a little girl. She liked candy.")
- Skeletal indication of time and place often held in past time ("once there was..."); little relationship to other narrative elements
- Beginning relationship between setting and other narrative elements (futuristic setting to accommodate aliens and spaceships); beginning symbolic functions of setting (often stereotypical images—forest as scary place)
- Setting becomes more essential to the development of the story in explicit ways: characters may remark on the setting or the time and place may be integral to the plot
- Setting may serve more than one function and the relationship between functions is more implicit and symbolic—for example, setting may be linked symbolically to character mood ("She hid in the grass, clutching the sharp, dry spikes, waiting.")
- Setting fully integrated with the characters, action, and theme of the story; role of setting is multifunctional—setting mood, revealing character and conflict, serving as metaphor



Plot

simple ⟷ complex
static ⟷ conflict

- One or two events with little or no conflict ("Once there was a cat. The cat liked milk.")
- Beginning sequence of events, but occasional out-of-sync occurrences; events without problem, problem without resolution, or little emotional response
- Single, linear episode with clear beginning, middle, and end; the episode contains four critical elements of problem, emotional response, action, and outcome
- Plot increases in complexity with more than one episode; each episode contains problem, emotional response, action, outcome; beginning relationship between episodes
- Stronger relationship between episodes (with the resolution in one leading to a problem in the next); beginning manipulation of the sequence through foreshadowing, and subplots
- Overarching problem and resolution supported by multiple, episodes; rich variety of techniques (building suspense, foreshadowing, flashbacks, denouement) to manipulate sequence

Communication

context-bound ⟷ reader-considerate
literal ⟷ symbolic

- Writing bound to context (You have to be there) and often dependent on drawing and talk to clarify the meaning; minimal style and tone
- Beginning awareness of reader considerations; straightforward style and tone focused on getting the information out; first attempts at dialogue begin
- Writer begins to make use of explanations and transitions ("because" and "so"); literal style centers on description ("sunny day"); tone explicit
- Increased information and explanation for the reader (linking ideas as well as episodes); words more carefully selected to suit the narrative's purpose (particularly through increased use of detail in imagery)
- Some experimentation with symbolism (particularly figurative language) which shows reader considerations on both explicit and implicit levels; style shows increasing variety (alliteration, word play, rhythm, etc.) and tone is more implicit
- Careful crafting of choices in story structure as well as vocabulary demonstrate considerate orchestration of all the available resources; judicious experimentation with variety of stylistic forms which are often symbolic in nature and illuminate the other narrative elements

Wolf/Gearhart/Quellmalz/Whittaker, 1993

Figure 1. WWYR narrative rubric.

Table 3

Comparison Progress Rubric

-1	Some decline in competence of narratives
0	No change in competence of narratives
1	Slight increase in competence of narratives
2	Moderate increase in competence of narratives
3	Marked, striking increase in competence of narratives
N/A	Can't score

Please note that we use the term "narrative collection" rather than "narrative portfolio" advisedly: The raters in this study were presented with sets of narratives that were not selected by students or teachers to demonstrate particular competencies, nor did these sets include supplementary sources of evidence, such as a student cover letter, student self-assessments of individual narratives, or teacher assessments. While definitions of "portfolio" do vary markedly, we recognize that a collection does not a portfolio make. Therefore, our study should be interpreted as an examination of some of the technical issues surrounding the scoring of multiple samples of writing.

NARRATIVE ASSESSMENT

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), and its use has been shown to impact teachers' knowledge and practice (Gearhart, Wolf, et al., 1994; Gearhart & Wolf, 1994). The rubric differs from most narrative rubrics in its narrative-specific content and its developmental framework (Figure 1).

A recent technical study provided evidence of the rubric's reliability and validity (Gearhart, Herman, et al., 1994). We evaluated the rubric against an established rubric (the same rubric used in Baker, Gearhart, & Herman, 1991; Gearhart et al., 1992; Herman et al., 1993) that has consistently demonstrated sound technical capabilities in large-scale use (Table 1). Our findings indicated that at least three of the six WWYR scales can be used reliably and meaningfully for scoring classroom narratives, provided that each narrative is

rated by two raters. There were several sources of evidence for the validity of the WWYR rubric: The scores from both rubrics produced a pattern of increasing competence with grade level; WWYR scores were highly correlated with the comparison scores; comparisons of raters' judgments made with both rubrics for the same narratives indicated some consistency in their decisions; raters felt that the content of the WWYR rubric captured more aspects of narrative than the comparison rubric and had greater instructional potential. Neither the WWYR nor the comparison rubric produced patterns of highly distinctive subscale judgments. Thus, although raters agreed that WWYR scales had greater instructional utility than comparison scales and that each of the WWYR scales had relevance for instructional planning and classroom assessment, the subscale judgments did not provide a technically sound profile of students' strengths and weaknesses.

The present study builds on our technical study of the WWYR narrative rubric in several ways. First, technical issues concerning assessment of single narratives are extended to the assessment of narrative collections. We developed a WWYR holistic scale that integrates key concepts from the original scales (WWYR Overall Effectiveness scale, Table 4), and we examined its potential to support quantitative judgments of narrative collections. Second, we 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 (Figures 2 and 3). Third, we compared indices of students' narrative assessment across three assessment contexts (direct assessment, classroom samples, and collections) and two rubrics.

Table 4

WWYR Effectiveness Scale for the Narrative Rubric

Narrative Rubric
Overall judgment: How are features integrated in this narrative?

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 4 (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 itme, 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 4 (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 multi-functional 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 example available.

ID	Overall Effectiveness *(1–6)*	Progress *(-1–3)*	Strength and Weakness	
			Strength: Theme Plot Character Setting Communication INTEGRATION Weakness: Theme Plot Character Setting Communication	Commendation: Recommendation:
			Strength: Theme Plot Character Setting Communication INTEGRATION Weakness: Theme Plot Character Setting Communication	Commendation: Recommendation:

Figure 2. Narrative collection rating sheet for WWYR.

ID	General Comp. *(1–6)*	Progress *(-1–3)*	Strength and Weakness	
			Strength: Focus/Org. Elab./Devel. Weakness: Focus/Org. Elab./Devel. Both similar	Commendation: Recommendation:
			Strength: Focus/Org. Elab./Devel. Weakness: Focus/Org. Elab./Devel. Both similar	Commendation: Recommendation:

Figure 3. Narrative collection rating sheet for comparison rubric.

OUR STUDY

Our study addressed a set of related questions regarding the technical quality of the WWYR rubric.

> *Reliability: Can the* Writing What You Read *rubric be applied to scoring of narrative collections with the same levels of rater agreement as an established narrative rubric?*

We selected a comparison narrative rubric that has consistently demonstrated excellent levels of rater agreement. Raters scored narrative collections with both rubrics (scoring design described below), and we compared indices of rubric reliability.

> *Can the* Writing What You Read *rubric be applied to narrative collections with the same levels of rater agreement previously reported for assessment of single narratives?*

Findings for WWYR scoring of narrative collections are compared with our prior findings for WWYR scoring of single narratives.

> *Validity of the* Writing What You Read *rubric: What is the evidence that collection scores derived from the WWYR rubric are or are not meaningful indices of students' narrative writing?*

We inferred validity from grade-level differences (scores should increase with age), from relationships of scores across rubrics (e.g., collection scores derived from both rubrics should be correlated), from relationships across assessment contexts (e.g., scores derived from a direct assessment, samples of classroom narrative writing, and narrative collections should be correlated), from consistency of raters' collection judgments across rubrics, and from raters' confidence in their collection judgments based on opinions expressed in post-rating interviews.

Procedures

Site

The narrative samples were collected from an elementary school that served as a longitudinal research site for the national Apple Classrooms of Tomorrow project. The school is located in a middle class suburb of Silicon Valley.

Datasets

There were three datasets: *direct assessments,* samples of *classroom narratives,* and *narrative collections.* Students' names and grade levels were removed from all material and replaced with identification numbers. Narratives or narrative collections were sorted by level (primary = Grades 1 and 2, middle = Grades 3 and 4, and upper = Grades 5 and 6) and then scrambled within sets.

The *direct assessments* were narratives written to prompts designed by the same school district that currently utilizes the comparison rubric. At the request of the teachers at our site, students in Grades 2 through 5 responded to a "magic" prompt (see Appendix A), and students in Grade 6 to a sports prompt (Appendix B). Following the procedures established by the comparison district, these performance-based assessments were administered over two days. On the first day, students discussed literatures 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.

The *classroom narratives* were sampled from assignments in Grades 2 through 6; content criteria for identifying writing as a narrative were established by Shelby Wolf (Wolf & Gearhart, 1993a, 1993b). For Grade 3 only, each of the narratives written by each of the students was scored separately to permit us to compare indices of narrative competence based on a mean of all class narrative assignments with the holistic score assigned each student's narrative collection.

The *narrative collections* were constructed by us from students' writing folders. While the folders of Grade 3 students contained all of their narratives (usually within a range of 3–6 narratives), the collections of all other students contained a sampling of 3–6 narratives sequenced by date. The limitation on number was designed to reduce the complexity of the scoring task, and for most students the range of 3–6 enabled us to include each narrative. Because students' folders varied in their inclusion of process materials (initial brainstorm, first draft, revisions), raters were provided only final drafts (or the last draft available). Table 4 shows the number of students from whom we had both on-demand assessments and classroom narratives.

Rating Procedures

Raters. Five experienced raters participated, all of whom had participated in our earlier technical study of the WWYR rubric (Gearhart, Herman, et al., 1994) and thus had considerable training and experience with both the WWYR and the comparison rubric. Two were elementary teachers with previous experience using the comparison rubric for scoring students' narrative writing; one of these raters had considerably more experience than the other with district scoring sessions. A third rater was an elementary teacher experienced with scoring elementary narrative and persuasive writing samples in English and Spanish for program evaluation. The fourth rater was a research assistant with experience scoring elementary narrative and persuasive writing samples in English and Spanish for program evaluation. Involved in this study only with the scoring of one dataset, the fifth rater was an elementary teacher with experience scoring elementary narrative and persuasive writing samples in English and Spanish for program evaluation.

Rating. This study and its companion were conducted in the summer of 1993 and on Saturdays as feasible during the 1993-94 year. Scheduling constraints impacted the design of the rating procedures. The sequence of scorings is listed below.

Summer 1993

Direct narrative assessments—comparison rubric
Classroom narratives—comparison rubric
Classroom narratives—WWYR rubric

Saturdays 1993-94

Narrative collections, first half—WWYR rubric
Narrative collections, second half—comparison rubric
Narrative collections, first half—comparison rubric
Narrative collections, second half—WWYR rubric
Direct narrative assessments, Grade 3 only—WWYR rubric

The order of scoring collections was designed to reduce possible interactions of order of scoring and rubric on raters' judgments.

At each scoring session, raters scored narratives 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 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 such that disagreement among raters on any scale was not greater than 0.5. (Raters requested permission to locate ratings at midpoints in addition to 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 conducted an additional training session; raters scored 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 that made certain that raters were not changing their criteria for scoring. 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. The material was distributed so that two raters rated each piece independently, scores were compared (by one of this report's authors), and a third rater rated any paper whose scores on any scale differed by more than one scale point.

When scoring the narrative collections, raters applied each rubric's holistic scale, made additional judgments regarding a collection's strength and weakness based on each rubric's analytic subscales (Figures 2 and 3), and then wrote brief commendations and recommendations.

Rater Reflection

A focus group interview was conducted at the completion of all scoring (findings from previous interviews are reported in Gearhart, Herman, et al., 1994). The interview were transcribed for analysis; the protocol is contained in Appendix A.

Results

Reliability of Narrative Collection Scores

Analysis of the reliability of the narrative collection scores was hampered by the small size of the dataset. The scoring proved to be quite challenging to the raters and proceeded more slowly than initially expected. As a result, only 52 collections were scored. Although the results indicate some interesting differences between rubrics in the patterns of reliability, the small sample size precludes us from making strong inferences about issues of relative reliability.

Interrater agreement. Four raters participated in the rating process. 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. 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.[1] For this study, percent agreements between rater pairs were computed based on exact, ±0.5, and ±1.0 criteria. Table 5 summarizes the results.

The statistics in Table 5 indicate both promise and problems. The promise is indicated by the consistently higher levels of agreement obtained for the WWYR rubric relative to the comparison rubric. The problem is indicated by the small sample sizes for each rater pair, ranging from 5 to 17. These small sample sizes 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 relative stability of ratings may be assessed through the use of correlations between rater pairs. While classical reliability coefficients are defined as the correlations between parallel forms of

[1] 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.

Table 5

Percent Agreement to Within Specified Criteria for Rater Pairs

Rubric	Rater Pair	*N*	Agreement criterion ±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

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

the same test, we can use them here to assess the stability of a student's scores across different (parallel) raters. Table 6 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. Those figures are .46 and .67, respectively. The reliability of the comparison rubric as measured by this index is well below what is desirable, while the WWYR rubric is very close to what would be considered adequate reliability for many purposes.

Table 6

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.

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

Generalizability of narrative collection scores. Variance components were estimated for a generalizability study with one facet, raters. This methodology enables us to identify the proportions of variability that can be attributed to collections (the universe of observation), raters (the single facet), and the interaction between collections and raters (error). Once variance components have been estimated, they may be used to construct generalizability coefficients appropriate to the purposes of the study. Since the data matrix was unbalanced to a large degree, two methods of estimating the variance components were used. The first, the MIVQUE method, is a closed form computational method that is easy to compute but may not be appropriate if the data are unbalanced. The second method, REML (REstricted Maximum Likelihood), is an iterative technique which is computationally expensive but may yield more stable estimates when data are unbalanced.

The results of these analyses are presented in Table 7. This table contains variance components for each rubric, as well as generalizability coefficients based on those variance components. Both relative and absolute coefficients are presented, the difference being that the absolute coefficients treat the rater as a source of error, while the relative coefficients ignore the variance due to raters. Absolute coefficients are most appropriate when scores are being compared to an absolute standard, such as a cut-score for a proficiency classification, while relative coefficients are called for when we are primarily

interested in the relative rankings of the scores. In addition, generalizability theory allows us to predict what the effect of adding more raters would be on reliability. This is illustrated in Table 7 as the generalizability coefficients for scores based on two raters.

The pattern of results here parallels those found in the earlier analyses. Once again, the WWYR rubric seems to perform better as a method for scoring narrative collections than 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 must repeat that these inferences are based on a very small sample size and a questionable design for this type of analysis.

Reliability of progress scores. In addition to the holistic scores assigned to each collection, raters also made judgments about the progress demonstrated by the writer over the course of the collection of the narratives. The analysis of the reliability of progress scores was hampered by the same sample size problems that affected the holistic scores, only to a greater degree. The already small number of collections scored was further diminished by the failure of the raters to assign progress scores to many of the collections. Table 8 contains correlations of progress scores for the rater pairs. Here we seem to see a reversal of the pattern of the holistic score analyses, with the comparison rubric outperforming the WWYR rubric. The mean unweighted correlation

Table 7

Results of the Generalizability Study for the Narrative Collection Scores

					Generalizability coefficients			
		Variance components			1 Rater		2 Raters	
Rubric	Method	Person	Rater	Error	Relative	Absolute	Relative	Absolute
Comparison	MIVQUE	0.3439	0.1145	0.4950	0.41	0.36	0.58	0.53
Comparison	REML	0.4895	0.1640	0.3699	0.57	0.48	0.73	0.65
WWYR	MIVQUE	0.3827	0.0206	0.1916	0.67	0.64	0.80	0.78
WWYR	REML	0.3619	0.0483	0.1930	0.65	0.60	0.79	0.75

Table 8

Correlations Between Rater Pairs for Narrative Collection Progress Scores

	Comparison				WWYR		
Rater	1	2	3	Rater	1	2	3
2	.17			2	-.29		
	(10)				(8)		
3	.91**	.40		3	-1.00	.59	
	(7)	(14)			(2)	(10)	
4	.43	.52	.07	4	.12	.67	.48
	(11)	(9)	(8)		(5)	(8)	(9)

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

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

for the comparison rubric is .42, and for the WWYR rubric .10. The weighted means (by cell size) are .40 and .28 respectively.

Reliability of Direct Assessments

Findings regarding the reliability of the direct assessment scores are reported here to provide the background necessary to interpret relationships of scores across assessment contexts—narrative collections, direct narrative assessments, and classroom narrative assignments. Note that the reliability of WWYR and comparison scorings of classroom narrative assignments was reported in Gearhart, Herman, et al., 1994.

Interrater correlations. Tables 9 and 10 report interrater correlations for direct assessments with each rubric. While there are ample data to make firm inferences about the comparison rubric, this is not the case for the WWYR rubric. To summarize the information in the tables, for the comparison rubric the weighted mean correlation is .68, and the weighted mean correlation for the WWYR rubric is .76. Although the WWYR rubric appears to outperform the comparison rubric, the small sample size renders this estimate less stable than desired.

Generalizability of direct assessments. Some technical problems arose in the estimation of variance components to be used in the computation of generalizability coefficients (Table 11). For the comparison rubric, there was a

Table 9

Interrater Correlations for Direct Assessment (DA) Scores on the Comparison Rubric (Each cell contains the correlation, the *p*-value, and cell *n*)

Rater	1	2	3	4	5
1	1.00 .000 *150*	.70 .000 *75*	.70 .000 *93*	-.11 .807 *7*	.70 .000 *84*
2	.70 .000 *75*	1.00 .000 *142*	.69 .000 *92*	.31 .297 *13*	.70 .000 *83*
3	.70 .000 *93*	.69 .000 *92*	1.00 .000 *157*	.58 .036 *13*	.70 .000 *85*
4	-.11 .807 *7*	.31 .297 *13*	.58 .036 *13*	1.00 .000 *31*	.77 .015 *9*
5	.70 .000 *84*	.70 .000 *83*	.70 .000 *85*	.77 .015 *9*	1.00 .000 *139*

Table 10

Interrater Correlations for Direct Assessment (DA) Scores on the WWYR Rubric (Each cell contains the correlation, the *p*-value, and cell *n*)

Rater	1	2	3	4
1	1.00 .000 *23*	.51 .089 *12*	.84 .001 *12*	.87 .000 *11*
2	.51 .089 *12*	1.00 .000 *24*	.85 .000 *13*	.70 .012 *12*
3	.84 .001 *12*	.85 .000 *13*	1.00 .000 *33*	.79 .001 *13*
4	.87 .000 *11*	.70 .012 *12*	.79 .001 *13*	1.00 .000 *23*

Table 11

Results of the Generalizability Study for the Direct Assessment Scores

					Generalizability coefficients			
		Variance components			1 Rater		2 Raters	
Rubric	Method	Person	Rater	Error	Relative	Absolute	Relative	Absolute
Comparison	MIVQUE	0.7364	0.0333	0.4426	0.62	0.61	0.77	0.76
Comparison	REML	0.8617	0.0436	0.3594	0.71	0.68	0.83	0.81
WWYR	MIVQUE	0.4117	0.1037	—	—	—	—	—
WWYR	REML	0.2639	0.0659	0.1007	0.72	0.61	0.84	0.76

Note. For the WWYR rubric the MIVQUE estimation method produced a negative variance estimate for the error, and so generalizability coefficients could not be computed.

considerable difference in the estimates obtained through the two estimation methods. We have presented both sets of estimates here, but we choose to interpret those obtained through the REML method in the belief that the iterative procedure will be less sensitive to the unbalanced nature of the data. The relative and absolute generalizability coefficients for the comparison rubric are .71 and .68, respectively, which are in the acceptable range. The corresponding coefficients for scores derived based on two raters' judgments are .83 and .81 and indicate a good level of reliability.

For the WWYR rubric, the MIVQUE method provided a negative estimate of the error variance.[2] Positive estimates were obtained through the use of the REML method, and so those estimates were used to compute generalizability coefficients. Nevertheless, the failure of the MIVQUE method to provide reasonable estimates is indicative of possible problems in the data. The pattern of generalizability coefficients for the WWYR rubric was quite similar to the

[2] Because variances are sum of squared deviations from the mean, and because squares are always positive, variances should always be positive quantities. However, the computational methods used by procedures like MIVQUE can sometimes produce estimates of these variances that are negative. There are two principle reasons why this could happen. First, all estimates are subject to sampling error, which is measured by the standard error of the estimate (the standard deviation of the sampling distribution). If the true value of the variance component is zero or close to zero, then the negative estimates may be within the range of what we might expect to be the expected variation. Another possibility is that the data is ill-conditioned in some way; e.g., perhaps some kind of an outlier case is unduly influencing the estimation process.

pattern obtained for the comparison rubric. We conclude that there is evidence that scores for direct assessments derived using both rubrics are reasonably reliable, especially if the scores of two raters are averaged.

Validity

This section contains analyses of the *Writing What You Read* rubric's capacity to produce meaningful results when applied to narrative collections: (a) comparisons of students' scores across grade levels (we would expect scores to increase with grade level), and (b) comparisons of scores across assessments and rubrics (we would expect raters to make similar judgments of the same assessment using either rubric, and, we would expect raters to make similar judgments of all assessments with one rubric). All ratings contributed to these results: Scores were computed as the mean of the independent ratings *or* the resolved score achieved through discussion during the training and check sets.

Grade level comparisons. Table 12 and Figure 4 contain descriptive statistics for each rubric-assessment combination. For each rubric and each assessment, there were score differences in the expected direction by grade level.

ANOVAs (Tables 13-18) were designed to contrast linear, quadratic, and cubic trends for each rubric/assessment combination, with the exception of WDA which only had two grade levels scored. In such modeling, the sum of squares that is attributable to grade level is further decomposed into portions representing linear, quadratic, cubic trends, or any combination of these trends or higher degree trends. A linear relationship between grade and performance would indicate that as grade increases, the scores increase, and that the rate of increase would be constant: 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, if scores initially increased, reached a peak, and then decreased, this could show up as a quadratic trend. More complex patterns can be modeled by cubic trends. As shown in Table 12, the linear trend was significant for all of the variables and provided the strongest contribution to the sum of squares due to grade level. For the CCLASS and CDA variables, both the cubic trend and a combined linear and cubic trend were also significant. As shown in the

Table 12

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) (Cells contain means, standard deviations, and *n*s)

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

CCLASS box-plot (Figure 4), for example, the scores make a sizable increase from Grade 2 to Grade 3, level off through Grade 4, and then increase dramatically again at Grade 6. Thus the descriptives and significance tests support the "developmental validity" of all measures.

It is, however, difficult to make any kind of comparison between the two rubrics just by looking at the patterns of means. Each of the means is an estimate, and each estimate is subject to sampling variability. That variability itself is quite variable across the rubric-assessment combinations and across the grade levels. To illustrate this aspect, we generated plots that showed the approximate 95% confidence intervals for the means of the rubric-assessment combinations across grade levels (Figure 5). We can see, for example, that the

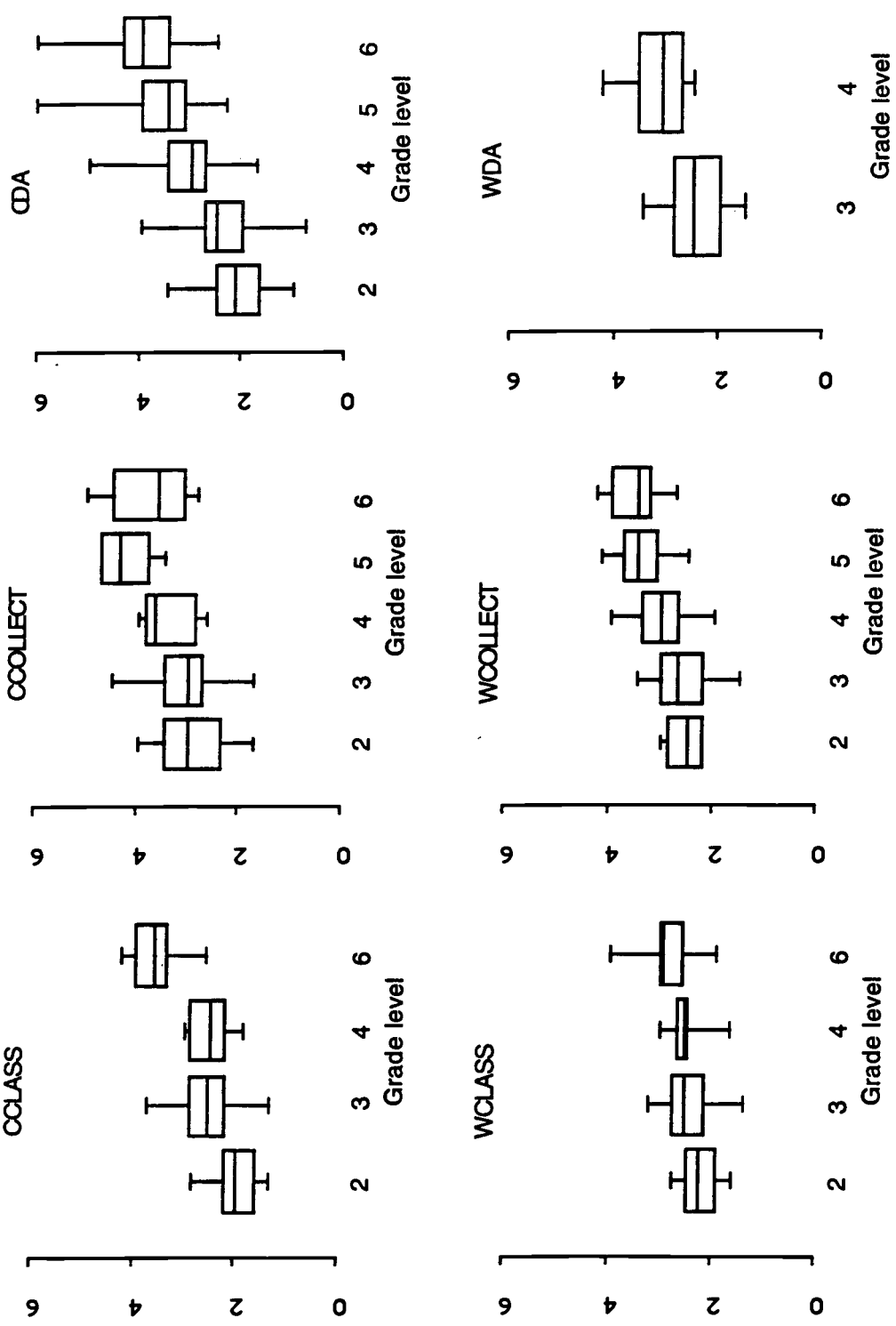


Figure 4. Boxplots showing the median, interquartile range, and score range for each rubric (Comparison C and *Writing What You Read* W) for three assessments (classroom narratives CLASS, narrative collections COLLECT, and direct assessments DA).

Table 13

ANOVA Table for Comparison Rubric Scores of Classroom Narratives (CCLASS)

Source	*DF*	Sum of squares	Mean square	*F* value	Pr > *F*
Grade	3	21.708	7.236	30.73	0.0001
Error	65	15.305	0.235		
Total	68	37.013			
Contrasts					
Linear	1	20.860	20.860	88.59	0.0001
Quadratic	1	0.003	0.003	0.01	0.9034
Cubic	1	1.278	1.278	5.43	0.0229
Linear-Cubic	2	21.700	10.850	46.08	0.0001

Table 14

ANOVA Table for Comparison Scores for Narrative Collections (CCOLLECT)

Source	*DF*	Sum of squares	Mean square	*F* value	Pr > *F*
Grade	4	12.580	3.145	6.48	0.0003
Error	48	23.280	0.485		
Total	52	35.860			
Contrasts					
Linear	1	9.074	9.074	18.71	0.0001
Quadratic	1	1.234	1.234	2.55	0.1172
Cubic	1	0.028	0.027	0.06	0.8129

Table 15

ANOVA table for Comparison Scores for Narrative Direct Assessments (CDA)

Source	*DF*	Sum of squares	Mean square	*F* value	Pr > *F*
Grade	4	118.675	29.669	64.94	0.0001
Error	238	108.729	0.457		
Total	242	227.404			
Contrasts					
Linear	1	116.412	116.412	254.82	0.0001
Quadratic	1	0.169	0.169	0.37	0.5441
Cubic	1	2.068	2.068	4.53	0.0344
Linear - Cubic	2	118.480	59.240	129.67	0.0001

Table 16
ANOVA Table for WWYR Scores for Classroom Narratives (WCLASS)

Source	*DF*	Sum of squares	Mean square	*F* value	Pr > *F*
Grade	3	3.033	1.011	4.68	0.0050
Error	65	14.026	0.216		
Total	68	17.059			
Contrasts					
Linear	1	2.950	2.950	13.67	0.0004
Quadratic	1	0.049	0.049	0.23	0.6366
Cubic	1	0.107	0.107	0.49	0.4848

Table 17
ANOVA Table for WWYR Scores for Narrative Collections (WCOLLECT)

Source	*DF*	Sum of squares	Mean square	*F* value	Pr > *F*
Grade	4	8.639	2.160	7.30	0.0001
Error	45	13.310	0.296		
Total	49	21.948			
Contrasts					
Linear	1	8.096	8.096	27.37	0.0001
Quadratic	1	0.067	0.067	0.23	0.6356
Cubic	1	0.334	0.334	1.13	0.2939

Table 18
ANOVA Table for WWYR Scores for Narrative Direct Assessments (WDA)

Source	*DF*	Sum of squares	Mean square	*F* value	Pr > *F*
Grade	1	7.308	7.308	26.62	0.0001
Error	60	16.474	0.275		
Total	61	23.782			

most stable and interpretable trend occurs for the CDA variable, or the direct assessments scored with the Comparison rubric. This stability is due largely to the relatively large sample sizes for that combination across all grade levels.

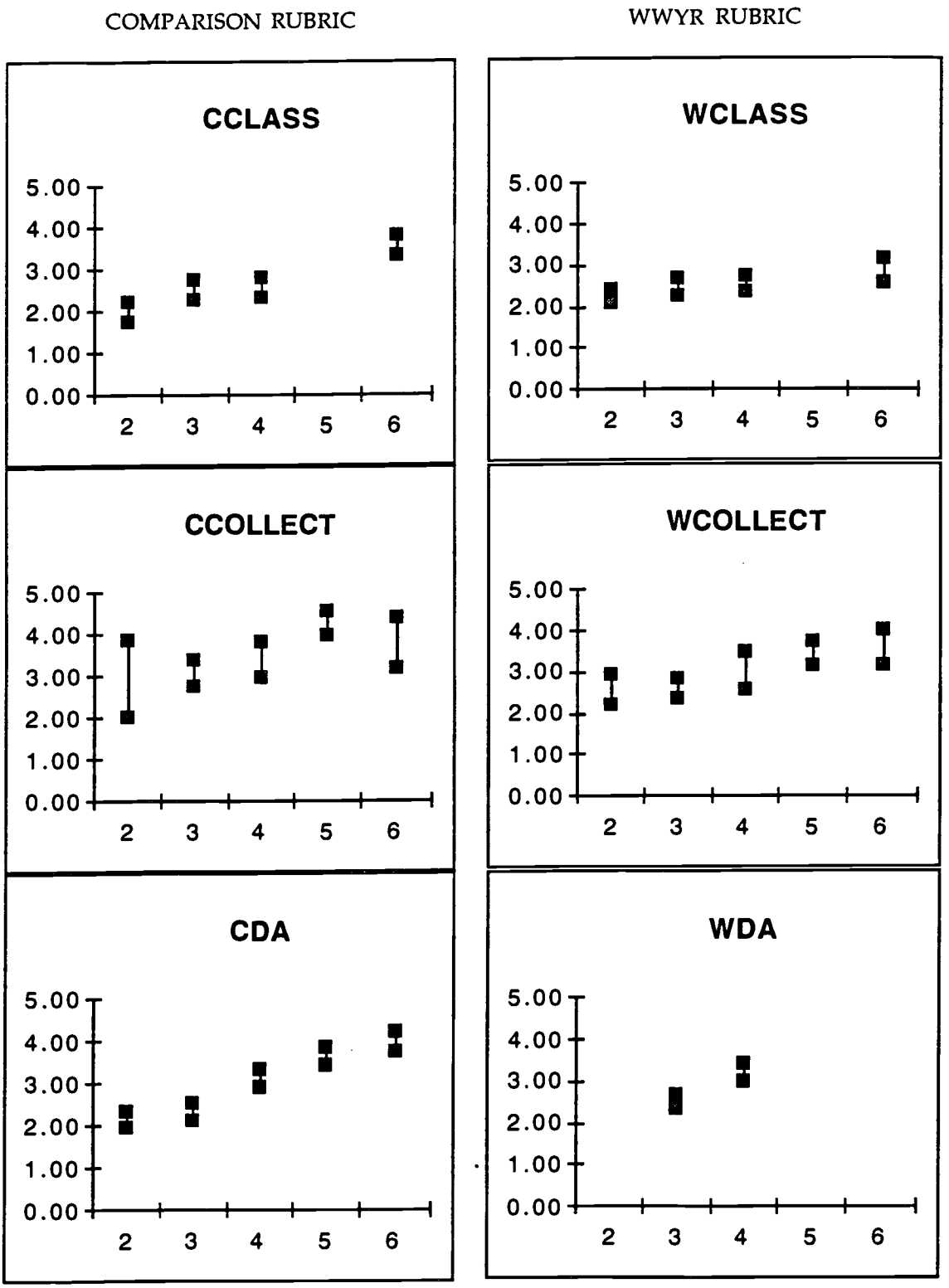


Figure 5. Plots of approximate 95% confidence intervals for the means of the rubric-assessment combinations plotted across grade levels.

In contrast, we see a great deal of instability for the narrative collections scored with the same rubric (CCOLLECT). We can attribute the inordinately wide confidence bands at the second- and fifth-grade levels there to the extremely small sample sizes (four and eight collections, respectively) for those combinations. We seem to see a more stable pattern for the same combination when scored by the WWYR rubric (WCOLLECT), even though the sample sizes are equally small; but again we are very reluctant to say more than this due to the small number of observations that we have to go on.

Comparisons across assessments and rubrics. We examined relationships across assessments (class assignments, narrative collections, and direct assessments) and rubrics (WWYR, comparison).

Means and variances. Table 19 presents descriptive statistics. Only scores from subjects in the third grade are included, because this grade level was the only level in which all of students' narrative work was scored. We see a remarkable degree of consistency with respect to means and standard deviations across the various rubric-scale combinations, with the sole exception of the comparison rubric as applied to the narrative collections (CCOLLECT). With respect to within scale-between rubric comparisons, only the CCOLLECT-WCOLLECT comparison is significant, with a repeated measures F statistic of 10.55 with 1 and 19 degrees of freedom ($p < .01$). The between-scale–within-rubric comparison for the comparison rubric was also significant, $F(2, 19) = 17.74$, $p < 0.000$, while the same comparison for the WWYR rubric was not significant.

Table 19

Descriptive Statistics for Two Rubrics Applied to Three Assessments, Scores From 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 20 contains all correlations across assessments and rubrics. In such a table we might expect to see the highest correlations when we have (a) the same assessment scored with different rubrics (the elements on the diagonal of the lower left quadrant of Table 20), and (b) different assessments 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 assessments are measured 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 scales as scored by the WWYR rubric are quite consistent and respectable, we see no such consistency between the scales as scored with the comparison rubric. In fact, the correlation between CCOLLECT and CDA is the smallest that appears 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, a large part of the explanation for these inconsistencies can perhaps be attributed to the anomalous performance of the comparison rubric as applied to the narrative collections.

Consistency of mastery decisions. One potential use of scoring rubrics is to make decisions about students' mastery of skills or competencies. Such a usage requires that some cutpoint for mastery first be chosen; 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. 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 21.

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

Table 20

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

Correlation *p*-value *N*	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*	

Table 21

Contingency Table for Examining Decision Consistency

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

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.

There are at least two important issues to consider relative to our application regarding the question of decisions made by using different rubrics. The first is the underlying assumption that both rubrics are really measuring the same latent trait. Even if this is true (and this is subject to empirical validation), there may be problems related to the second issue, that is, the setting of appropriate cutpoints. Consider a case in which two methods are perfect indicators of some latent trait, to the degree that the distributions of the manifestations of this latent trait as translated into method scores 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, there is no way that we could obtain good decision consistency. Yet 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. 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.

The implication of the above discussion is that if we are interested in equating methods of making mastery decisions, then careful attention must be paid to both the underlying factor structure of the measurements, and to the distributions of the scale scores. The appropriate methodology for addressing these questions may be confirmatory factor analysis (CFA), first as a method of examining the assumption of unidimensionality, and second as a means of transforming the observed scores, perhaps into factor scores, so as to

standardize their distributions and facilitate the setting of appropriate cutpoints. Unfortunately, we again find ourselves in the regrettable position of not being able to pursue such an approach due to our lack of data. We do have some empirical evidence for unidimensionality across some of the scales and rubrics as provided by the correlations between the various scores, and, with the exception of the CCOLLECT scale, it appears that the distributions of the scores on the different rubric-assessment combinations are reasonably similar. Again, however, due to the small amount of data presented, the reader should exercise caution (as we will) in the degree of belief which s/he holds regarding our findings.

Measuring decision consistency. The most commonly used statistic for measuring decision consistency is the kappa coefficient. 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 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.

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 sampling distribution had a standard deviation of .29 and was noticeably skewed in the negative direction. Figure 6 is a histogram of this sampling distribution.

Decision consistency across the narrative scales and rubrics. 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 commonly used (e.g., Gearhart, Herman, et al., 1994). Although the conditions

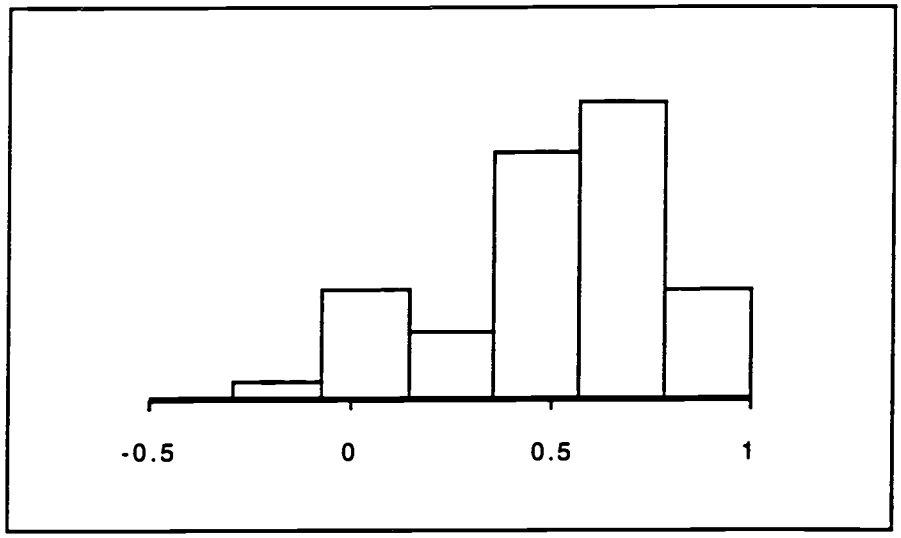


Figure 6. Histogram of the distribution of kappa coefficients from 100 iterations of a simulation process.

for this study differ considerably from prior usage of the rubric, it was decided to start with that cutpoint. The WWYR rubric has little history as a measure of narrative competency, and so, given the lack of prior experience and the observed similarities in the distributions of the various rubric-assessment combinations, it was decided to begin with the same cutpoint for that rubric. Table 22 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.

Examination of this table shows us 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.

Table 22

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		

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. This is a very clear amplification of a situation in which the statistic of choice exhibits a degree of stability that falls far short of what is desirable.

Adjustment of the cutpoints. This lack of stability is primarily a function of the cutpoints, which appear to set a standard for performance that is unattainable by most of the sample. Given the developmental nature of the rubric contents and the early developmental stature of the subjects (Grade 3), it was decided to adjust the cutpoints downward. Table 23 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 by the simulation study; that is, we are getting decision consistencies that are consistent with what we would expect based on adequately reliable measures with appropriately comparable cutpoints.

The major exceptions to the findings above are found in those rubric-assessment pairs involving the comparison scores of the narrative collections (CCOLLECT). This is not surprising given the discussion earlier regarding the necessity of having comparable distributions for the measures, coupled with the observation 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 a clear indication that the comparison rubric as applied to the narrative collections functions somewhat differently than the other rubric-assessment combinations. The obvious solution would be to adjust the cutpoint upward for CCOLLECT, but even this did not provide an adequate solution. While the kappa coefficients for CCOLLECT with the other measures were .17, .34, .34, .17, and .29 using the common 3.0 cutpoint, adjusting the cutpoint upwards to 3.5 for the CCOLLECT scale resulted in respective kappa's of .21, .12, .12, .15, and -.32, or a significant deterioration of the statistics.

Table 23

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		

Summary. It would appear from the results presented above (repeating the caveats regarding the small sample sizes) that 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 scores for the narrative collections. As to why this measure differed from the others, we can offer no explanation here, and this may be a likely topic for further investigation.

Qualitative Judgments of Collection Strength and Weakness

For each narrative collection, raters were asked to choose a rubric subscale that reflected their judgments of a collection's strength or a collection's weakness. The task was included to explore the capacity of either rubric to represent the competence of a narrative collection beyond a single holistic score. Unfortunately, the results reflected the exploratory nature of this component of the rating process: Whether rating with WWYR or the comparison rubric, raters were not likely to agree (Table 24). Although there were differences among rater pairs in patterns of agreement, there were very few pairs that reached adequate agreement on strength or weakness using either rubric. Note that although the results (for Strengths) appear slightly better for the comparison rubric, the lesser number of scales enhanced the likelihood of agreement.

There was evidence that raters were sometimes using the two rubrics to capture essentially the same characteristics of the narrative collections. Thus when identifying a collection's weakness, raters tended to identify plot when using the WWYR rubric and Focus/Organization when using the comparison rubric; when identifying a collection's strength, they tended to choose character or communication when using the WWYR rubric and development/elaboration when using the comparison rubric (Tables 25 and 26). These patterns reflect content overlaps across the WWYR and the comparison schemes.

Table 24

Rater Agreement on Collection Strengths and Weaknesses

Rubric and rater pair	Percent of agreement		Number of judgments
	Strengths	Weaknesses	
WWYR			
1–2	.36	.27	11
1–3	.40	.80	5
1–4	.60	.40	5
2–3	.08	.42	12
2–4	.54	.63	11
3–4	.10	.20	10
Weighted average	.31	.42	54
Comparison			
1–2	.36	.45	11
1–3	.38	.25	8
1–4	.44	.77	9
2–3	.46	.46	13
2–4	.54	.36	11
3–4	.75	.38	8
Weighted average	.48	.45	60

Table 25

Frequency of Choice of WWYR Subscale for Collection Strengths and Weaknesses

	Subscale					
	Overall	Theme	Character	Setting	Plot	Communi-cation
Strength	10	19	26	4	15	24
Weakness	2	10	18	6	41	19

Table 26

Frequency of Rater Choice of Comparison Subscale for Collection Strengths and Weaknesses

	Subscale		
	Focus/ Organization	Development/ Elaboration	Both areas
Strength	17	29	27
Weakness	23	17	30

Note. "Both areas" indicates judgments that the collection exhibited strengths and weaknesses in both components of the rubric.

Commendations and recommendations. Raters were asked to comment on the collection using the constructs and language of the rubric in current use. Although the raters decided to build their comments on their choices of strength or weakness—to write a commendation on the strength, a recommendation on the weakness—many raters wrote additional comments derived from additional scales. We classified the comments by their reference to one or more analytic scales.[3] As shown in Table 27, raters' WWYR comments were likely to differ in scale content, while the Comparison comments were more likely to share scale content. Note again, however, that the Comparison rubric's lesser number of scales favored agreement.

Like the judgments of strength and weakness, raters' comments clustered around certain subscales (Tables 28 and 29). Using WWYR, raters commented more often on communication, character, and setting, and—reflecting their choices of collection strength and weakness—there was a tendency to offer commendations on communication or character and recommendations on plot. Using the comparison rubric, raters' comments tended to contain content from both subscales, and commendations more often focused on development and elaboration, while recommendations focused on focus/organization. Raters' WWYR commendations were differentiated across more scales than the focus just on development/elaboration captured by

[3] All comments were classified by an assistant, and the resulting codings were then carefully reviewed by one of the authors. There were few instances of disagreement, and formal evaluation of coder reliability was not deemed necessary in the context of data that are here offered as illustrative.

Table 27

Content Analyses of Commendations and Recommendations: Frequency of Rater Agreement on Collection Strengths and Weaknesses (*N* shown in parentheses)

	WWYR		Comparison	
Rater Pairs	Commendation	Recommendation	Commendation	Recommendation
1–2	.67 (9)	.56 (9)	.82 (11)	1.00 (10)
1–3	.78 (9)	.56 (9)	.89 (9)	.75 (8)
1–4	.33 (12)	.58 (12)	1.00 (8)	.75 (8)
2–3	.38 (8)	.62 (8)	.89 (9)	.67 (9)
2–4	.67 (6)	1.00 (6)	.69 (13)	.83 (12)
3–4	.40 (10)	.40 (10)	.82 (11)	.67 (9)
Weighted average	.52 (54)	.59 (54)	.88 (61)	.78 (56)

Table 28

Content Analyses of Commendations and Recommendations: Frequency of Choice of WWYR Subscale for Commendations and Recommendations

	Subscale					
	Theme	Character	Setting	Plot	Communi-cation	Other
Commendation	11	33	8	24	39	4
Recommendation	13	24	9	41	21	4

Note. Raters' comments often contained content from more than one WWYR subscale.

Table 29

Content Analyses of Commendations and Recommendations: Frequency of Rater Choice of Comparison Scale for Commendations and Recommendations

	Subscale	
	Focus/ Organization	Development/ Elaboration
Commendation	44	57
Recommendation	47	43

Note. Raters' comments often contained content from both Comparison subscales.

the comparison commendations. Recommendations made using both rubrics were generally consistent with one another and mirrored the findings for choice of weakness: Raters were recommending greater coherence, motivation, and development (or "focus") of plots.

Usefulness of the comments. The raters were encouraged to devise their own approach to commentary as long as they built their comments on the rubric in current use. No rater had prior experience with written commentary, and perhaps for this reason, many of the comments seemed to us to have little potential to guide teachers or students. We asked a graduate student and former elementary teacher to read through all of the comments and sort them into those that he felt were potentially useful, and those that were not. His sort produced clusters that had distinctive characteristics. For example, weak comments included those that were brief, contained quotes of isolated rubric phrases, and/or made no reference to the text of any narrative in a student's portfolio. In contrast, stronger comments included those that were longer, more diverse in content (thus reflecting the rater's own views), included paraphrases of the rubric, and/or made reference to a student's narrative text (e.g., "Wonderful vivid detail—events of 'popcorn factory' "). Examples of weak and strong recommendations are provided in Figure 7 for illustration.

STRONG RECOMMENDATIONS	WEAK RECOMMENDATIONS
Comparison rubric No attempt to limit topic. Jumps around in story, hard to follow. Digressions and over elaboration interferes with reading. Needs clearer topic with a more clear sense of beginning, middle and end. Plots were not complete: little sense of beginning/end. No sequential development. Some events not logical or linked to other events. No clear sense of beginning/middle/end. Some events not logical.	Plots uneven. Overelaboration/digression. Elements not evenly developed. Some may be omitted. Vary transitions. More transitions with less digression. Needs elaboration. Possible slight digression. Events need even development. Details are sometimes unevenly placed.
Writing What You Read Settings need to be developed, they could potentially play an important role in the stories, yet are only alluded to. More description of time and place with more relationship between characters and plot. Little or no indication of what motivates the character. Characters need to be described; develop character feelings. You need to be aware of your reader. Move to a higher level of plot: try some foreshadowing and stronger relationship between episodes. Develop theme by moving away from explicitly stating the meaning in your story.	Too little rounding of characters. Could develop more in story Should move theme and plot to more complex levels. Develop events evenly. Characters could use more rounding. Continue more rounding in description/feeling.

Figure 7. Examples of weak and strong recommendations.

Summary. There were two principal findings. First, raters were not likely to agree either in their choices of a rubric scale to represent the strengths and weaknesses of narrative collections or in their written commentaries. This pattern was more pronounced for WWYR. Although the lesser number of subscales in the comparison rubric (two compared with five for WWYR) clearly favored agreement, the content of the WWYR subscales may also have supported "false disagreements": WWYR's subscales were designed to be distinctive yet highlight the integrated nature of narrative components (Wolf & Gearhart, 1993a, 1993b), and therefore a rater may have found herself forced to choose between two or three subscales to describe what may have been understood as essentially the "same" strength or weakness. Thus, for large-scale assessment, restricting the number of choices for strength and weakness and defining them distinctively may benefit clarity of judgments. Second, using either rubric, raters tended to select one rubric scale for strength and a different scale for weakness. Whether these patterns reflected characteristics of the narrative collections, biases among the raters, or both will require further investigation.

Raters' Reflections

We have previously reported the raters' critiques of both narrative rubrics for scoring single narrative assignments for either large-scale or classroom assessment (Gearhart, Herman, et al., 1994). Raters repeated many of the same concerns for the scoring of narrative collections. In brief, for large-scale use, they felt that the comparison rubric relied too much on comparative terms ("more," "less"), emphasized the existence of certain features without an evaluation of their use in the narrative, and, with only two scales, was limited in its capacity to capture the narrative genre; on the other hand, they felt that WWYR was very complex and challenging to use, and that two of the five scales (plot and theme) had content weaknesses. For classroom use, they felt that WWYR provided far more explicit guidance in the design of narrative instruction and assessment, but that the comparison rubric's focus/organization scale captured something important about competent narrative writing missing from WWYR.

We highlight here just the new issues that raters raised regarding the scoring of narrative *collections*.

Holistic scores. Following extensive experience with analytic scoring of single pieces, raters had some difficulty utilizing just the holistic score: "You couldn't capture . . . exactly what it was" about the child's narrative writing. They were therefore appreciative of the opportunities to use the analytic scales for strength and weakness, and to write commentary.

One score for a diversity of material. No rater had prior experience scoring portfolios or collections of writing, and they were challenged by the diversity of narrative material.

> It is difficult to assign a score when you're dealing with so many different types of narratives at one time . . . dealing with so many different elements [is a problem], and one essay might have certain things than another doesn't, so how do you put that in one general score?

Rubric interacted with the challenges of assessing a collection. One rater mentioned that scoring a diversity of narratives with WWYR was more difficult, and another commented that scoring with the comparison rubric was less difficult.

> WWYR is more difficult to use if you are evaluating more work, and you're trying to hold all those ideas in your mind and consider each paper when you're looking at the whole entire portfolio.

> I think that when you're evaluating a lot of work, that . . . it's just real clear [with the comparison rubric] as to where they are just overall, whether they are "emerging," "developing," or later. . . . If you want to decide what area to delve into, you can [use the scales], say, to teach them better to stay on topic, or, if you notice that development's the area of weakness, then you would work on development.

Biased selection of WWYR subscales for qualitative judgments and commentary. One rater summarized the strategy that she and other raters used for picking WWYR subscales for strength and weakness. She first read a piece to see if a "particular area jumped out at you," and then, if not, she went through the five subscales in a particular order: "plot, communication, theme, and then I would go to character and setting." Her order did not correspond to the left-to-right sequence of subscales on the handout (theme, character, setting, plot, communication), and thus the order reflected her perspective on what components of narrative are most critical to narrative competence. Although not every rater mentioned use of this strategy (indeed,

one rater replied, "I really used the benchmarks and compared them"), her description matches closely with the scale components most frequently addressed by raters in their qualitative judgments and commentary and thus helps to explain those findings.

Progress scores. The raters' discussion of both progress rubrics revealed how uncertain they were when scoring for progress. They pointed to the inclusion only of comparative criteria—for example, "slight," "moderate," "marked." With so little guidance, two raters observed that WWYR at least gave them "more specific ideas about what to look for," and indeed one rater noticed that, as a result, she tended to give a higher Progress score with WWYR than with the comparison rubric. Two raters felt that the progress score was redundant with the overall score, suggesting that they had not clearly differentiated the functions of the two scales.

Summary. Raters' comments reflected the challenges of rating collections of diverse material, and, in that context, there was mention of greater difficulty of utilizing a more complex rubric (WWYR) for scoring a more complex assessment (collections vs. single narratives). Raters' discussion of their approaches to our exploratory assessments—scoring progress, selecting a scale to represent a collection's strength and weakness, and writing commentary—confirmed that these remain approaches that will require further investigation. Raters will need more detailed rubrics, scoring procedures, and exemplars, and their assessments will need validation against other measures.

SUMMARY AND DISCUSSION

This report addressed technical questions regarding the reliability and validity of large-scale portfolio assessment: (a) Can raters score collections of narrative writing reliably with rubrics designed for scoring single samples? (b) How do two rubrics derived from different frameworks differ in their capacities to support technically sound assessments of narrative collections? (c) Do ratings of distinctive narrative assessments (e.g., direct writing assessment vs. collection of classroom writing) categorize groups similarly? Based on raters' judgments of only 52 collections of students' narrative writing, this study was principally designed to illustrate analytic techniques for addressing each of these questions. In addition, as feasible within the

context of a very small dataset, we sought evidence for the utility of our new *Writing What You Read* narrative rubric for large-scale assessment of narrative collections. The performance of the WWYR rubric was evaluated against a comparison rubric that has consistently demonstrated sound technical capabilities in large-scale use.

Raters scored collections of narratives with the holistic scales of two holistic/analytic rubrics. 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 studies. 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. An exploratory scale for assessing evidence of progress within the narrative collections was not found to be reliable.

There was mixed evidence of validity for the narrative collection scores. 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 (direct assessment, and the mean of students' individually-scored classroom narratives). 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. Thus, overall, the quantitative results favored the WWYR rubric. However, two sets of results raise questions about the meaningfulness of the WWYR results: first, unexplained *positive* relationships between the WWYR narrative collection scores and each of the Comparison measures; and second, the absence of raters' agreement on the strengths and weaknesses of the collections, as we discuss next.

In addition to holistic scoring of the collections, raters used the analytic scales of each rubric as a framework for two additional assessments: the choice of rubric scales to represent the strength and the weakness of each collection, and written commentary on the collection. For each of these exploratory approaches to collection assessment, raters were not likely to

agree. Although this pattern of disagreement was more pronounced for WWYR, the lesser number of scales in the comparison rubric (two compared with five for WWYR) clearly favored agreement. These findings suggest revisions of our methods for large-scale assessment to support greater clarity of judgment—reducing the number of choices for strength and weakness, and defining them distinctively.

Thus, illustrating a multimethod approach to the technical study of new performance assessments, our study has produced preliminary evidence that the holistic scale of the *Writing What You Read* narrative rubric—a rubric designed to enhance teachers' understandings of narrative and to inform instruction—can be used reliably and meaningfully in large-scale assessment of narrative collections. We did not find support for our exploratory assessments of narrative progress, collection strength, and collection weakness, and therefore further research and development are needed to find ways to supplement quantitative holistic scoring with analytic judgments that provide instructional guidance. The pursuit of large-scale writing assessments that can guide the work of teachers in the classroom is important. When rubrics are designed to capture qualities of distinctive writing genre, then they have greater potential to support teachers' professional development, opportunities to learn in the classroom, and substantive interactions in moderation sessions.

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.

Gearhart, M., & Herman, J. L. (in press). Building bridges across assessment contexts: Issues in the use of classroom writing for large-scale portfolio assessment. *Evaluation Comment.*

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. L., Baker, E. L., & Whittaker, A. K. (1993). *Whose work is it? A question for the validity of large-scale portfolio assessment* (CSE Tech. Rep. No. 363). 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., & Wolf, S. A. (1994). Engaging teachers in assessment of their students' writing: The role of subject matter knowledge. *Assessing Writing, 1*(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., & 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.

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.

APPENDIX A: Magic Prompt Text

Grades 3 and 5 Writing – Prompt #27

Writing Situation

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. They might be used in other ways, too. 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?

(If you do not want to write about yourself, you may make up a character.)

Writing Directions

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.

Include details that will let the reader see the situation, the characters, and the events that happened.

* Cluster * Pre-write * Brainstorm * Doodle *

List * Etc. * Out of space? Use the back.

APPENDIX B: Sports Prompt Text

1993 Grade 6 Writing – Prompt #23

Writing Situation

At some time everyone gets involved in sports. It can be a group sport such as softball, soccer or baseball. Or it can be an individual sport such as swimming, ping-pong, or hiking. Imagine a special situation that took place when you were playing a sport. It can be a happy or frustrating experience. It can be a situation that has happened or could have happened. It may have been at home, at school, or some other place. Think about the people involved, what happened, and how you felt while you were playing.

(If you do not want to write about yourself, you may make up a character.)

Write a story that tells about the special situation when you (or the character) were playing a sport. Help the reader to understand what made the situation special, how you (or the character) felt and acted while playing, and why. Be certain to let the reader know how the situation ended.

Writing Directions

Write a story that tells about the special situation when you (or the character) were playing a sport. Help the reader to understand what made the situation special, how you (or the character) felt and acted while playing, and why. Be certain to let the reader know how the situation ended.

Include details that will let the reader see the situation, the characters, and the events that happened.

* Cluster * Pre-write * Brainstorm * Doodle *

List * Etc. * Out of space? Use the back.

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement (OERI)
Educational Resources Information Center (ERIC)

NOTICE

REPRODUCTION BASIS

☐ This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

☒ This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").