Gearhart, M., Herman, J. L., Novak, J. R., & Wolf, S. A. (1995). Toward the instructional utility of large-scale writing assessment: Validation of a new narrative rubric. _Assessing Writing, 2 (2), 207-242.

ASSESSING WRITING, 2 (2), 207–242, (1995)

Toward the Instructional Utility of Large-Scale Writing Assessment: Validation of a New Narrative Rubric

Maryl Gearhart, Joan L. Herman,
John R. Novak, and Shelby A. Wolf
University of California, Los Angeles

In the press to design performance-based writing assessments to serve both policy and practice, methods for the large-scale assessment of student writing have undergone considerable scrutiny and revision. A key question concerns instructional utility: What kinds of writing assessments can serve at once both policy decisions *and* the needs of teachers and students in the classroom? Direct writing assessments have demonstrated the kinds of technical capabilities needed for large-scale assessment (e.g., Huot, 1990a, 1990b, 1993), but their constraints on topic, resources, and time allowed are limits on their relevance to good instructional practice (Freedman, 1991; Huot, 1993; Williamson, 1994). Portfolio assessments hold promise as supports for students' growth as writers—through engaging curriculum and the deep integration of reading with writing (Camp, 1993; Hewitt, 1993; LeMahieu, Eresh, & Wallace, 1992; Mills, 1989; Murphy, 1994; O'Neil, 1992, 1993; Saylor & Overton, 1993; Simmons & Resnick, 1993; Spalding, 1995; Vermont Department of Education, 1991)—but their technical quality is under-researched (Condon & Hamp-Lyons, 1994; Elbow, 1994; Gearhart & Herman, 1995; Herman & Winters, 1994; Herman, Gearhart, & Asch-

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.

This article was a collaboration among its authors. Maryl Gearhart and Joan L. Herman had responsibility for design and conduct of the study; John R. Novak for design, conduct, and write-up of the quantitative analyses; Shelby A. Wolf for design of the WWYR rubric and training of raters. We thank our raters—Kathy Beneiof, Virginia Espinosa, Rosa Valdes, Maria Vega-Vidales, and Kim Uebelhardt; Jamal Abedi and Jia Moody for statistical advice and assistance; and the teachers, students, and parents who permitted us access to the students' narratives.

Correspondence and requests for reprints should be sent to Maryl Gearhart, Center for the Study of Evaluation, University of California, Los Angeles, CA 90095.

bacher, in press; Huot, 1994; White, 1994). Clearly, if we want the results of writing assessments to be useful in the classroom, research is needed that identifies possible strategies for building bridges between the requirements of large-scale assessments and the needs of young writers.

In this article, we address issues regarding the possible disjunct between what is good for large-scale assessment and what is good for teaching and learning. Our study represents one attempt to "marry" the large-scale and classroom perspectives; it is the story of our effort to move an assessment framework out of the classroom into the large-scale context. We begin by presenting background and rationale for a new narrative rubric that was initially designed to support classroom instruction. We then clarify the technical qualities necessary for the rubric's use for large-scale assessment purposes and examine the rubric's performance in relation to these qualities. Our goals are to contribute new understandings regarding tensions between the utility and the technical quality of writing assessments, as well as a model for conducting this kind of research.

COORDINATING CLASSROOM AND LARGE-SCALE ASSESSMENT

Let's consider for a moment how different may be the purposes and methods of classroom and large-scale assessment. In the classroom, teachers are invested in assessing the writing competencies of their students—students with whom they are engaged and know well from daily classroom contact. There are numerous opportunities to build and confirm judgments of students' current capabilities, and teachers are not particularly concerned with comparing the abilities of students in their own class with students in other classes, or from other years. Teachers' primary motivation is to obtain information that can be used to improve their curriculum and methods of instruction, and they feel their needs are best filled by qualitative approaches to writing assessments, some adapted or developed by themselves. Outside of the classroom, in contrast, policymakers and the public demand "report cards" that indicate whether or not the instruction which they are paying for is effective. These stakeholders are interested not in particular students, but in the relative performance of groups of students, and in changes in performance of these groups over time: How are the students in their local educational context performing compared to other students in the district, state, nation, or even world?

If we frame this contrast in accordance with modern notions of validity (Messick, 1993), we see a situation in which the validity of an assessment methodology may be perceived quite differently by these two groups of stakeholders. This conflict arises out of the contrasting purposes intended by each group. An assessment that might be considered quite valid for purposes of informing day-to-day instruction might lack validity when viewed as an indicator to inform policymakers. On the one hand, a complex and multidimensional rubric could yield an accurate and informative picture of the proficiency of individual students, yet may fail to meet generally accepted technical criteria for large-scale use (e.g., lack of reliability), or may simply be too costly to implement on a large scale. On the other hand, an assessment that can yield highly reliable, easily interpretable, and affordable results for groups of students (i.e., matrix sampling as implemented by National Assessment of Educational Progress, Johnson & Carlson, 1994) may utterly fail to capture the performance of an individual student and thus lack utility for the classroom teacher.

Our contrast polarizes, of course, what one might mean by "classroom" and "large-scale" assessment, and thus it bypasses attention to the many approaches to negotiated assessment that are evolving at school and departmental levels (e.g., Broad, 1994; Condon & Hamp-Lyons, 1994; Elbow, 1994; Haswell & Wyche-Smith, 1994; Moss, 1992, 1994; White, 1994). The exaggerated comparison serves our purposes here, however, for our focus is on large-scale assessment for public comparison and high-stakes decisions—uses of assessments that necessitate attention to technical measurement issues. Working with available evidence that the forms and contents of large-scale assessment are likely to drive instructional practice (Herman & Golan, 1991; Shepard, 1991; Smith, 1991), we have been seeking assessments that have technical quality yet retain the qualitative characteristics that make them desirable targets for performance and provide teachers with useful information for instructional decisions. The design of writing rubrics and methods for scoring becomes a special case set within these broader issues.

DESIGNING RUBRICS FOR WRITING ASSESSMENT

Although there exists critical debate surrounding the assignment of quantitative scores as an appropriate strategy for large-scale writing assessment (Broad, 1994; Elbow, 1994; Haswell & Wyche-Smith, 1994), the scoring of writing is viewed by many as a method that has the potential to represent both valid judgments as well as valued dimensions of writing competence. Interests in instructional value have highlighted the importance of rubric content and structure. Ratings from rubrics whose scales or scale-point criteria are vague, confusing, or inconsistent with what is known about well-constructed and effective text are neither valid measures of the important qualities of good writing nor useful supports for effective instruction (Baxter, Glaser, & Raghavan, in press; Paul, 1993; Resnick, Resnick, & DeStefano, 1993; Wiggins, 1993, 1994; D.P. Wolf, 1993). In order to communicate to teachers, students, and others what is important in writing performance, writing rubrics must be derived from current English/language arts frameworks and reflect those analyses of the contents, purposes, and complexities of text.

WRITING WHAT YOU READ

The goal of this study was to produce evidence of validity for a rubric grounded in current literacy frameworks and whose use has been shown to enhance instructional practice, but whose technical quality is unknown. Our effort began in 1991 with school-based research on the role of teachers' interpretive assessments in guiding the growth of young writers, focusing on methods of conferencing and written commentary (S.A. Wolf & Gearhart, 1993a, 1993b). One product of this phase of work was the analytic, multileveled *Writing What You Read (WWYR)* framework to support teachers' interpretations of both the development of a piece of narrative writing and the development of their students as writers (S.A. Wolf & Gearhart, 1993a, 1993b; S.A. Wolf & Gearhart, 1994). While referred to as a "rubric," the WWYR framework (Figure 1) featured intentionally *un*-numbered labels to emphasize the value of qualitative assessment of a child's achievement within particular contexts. We documented the ways that elementary teachers used WWYR as a resource for qualitative analysis of students' writing and the resulting growth in their understandings of narrative and children's narrative (Gearhart & Wolf, 1994; Gearhart, Wolf, Burkey, & Whittaker, 1994; S.A. Wolf & Gearhart, 1995).

The WWYR framework contains five analytic dimensions for Theme, Character, Setting, Plot, and Communication (Figure 1), and a sixth, holistic assessment of a narrative's Overall Effectiveness constructed specifically for this technical study (Figure 2). Each dimension contains six levels designed to match current understandings of children's narrative development. The technical language of narrative is integral to WWYR, unlike the descriptors of many narrative rubrics that are not unique to narrative genre. Thus, words like topic (rather than theme), event (rather than episode), and diction (rather than style) create a sense of "genre generality" (Gearhart et al., 1994). The typical rubric scale for "organization" may not capture the orchestration of narrative components, and a scale for "development" may not capture the communicative aspects of style and tone that center on creating images—using language purposefully, metaphorically, and rhythmically to take the reader off the page and into another world. The typical focus on the narrative components of character, setting, and plot omits theme—the heart of narrative, a comment about life which illuminates the emotional content of the human condition.

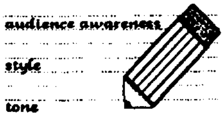
Theme	Character
explicit ◄──► implicit didactic ◄──► revealing	flat ◄──► round static ◄──► dynamic
• Not present or not developed through other narrative elements	• One or two flat, static characters with little relationship between characters; either objective (action speaks for itself) or first person (author as "I") point of view
• 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!")	• Some rounding, usually in physical description; relationship between characters is action-driven; objective point of view is common
• 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	• 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 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 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)
• 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	• 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)
• 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.)	• 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

Figure 1. Narrative rubric.

Action/Emotion
Over Time

Setting	Plot
backdrop ◄──► essential simple ◄──► multi-functional	simple ◄──► complex static ◄──► conflict
• Backdrop setting with little or no indication of time and place ("There was a little girl. She liked candy.")	• One or two events with little or no conflict ("Once there was a cat. The cat liked milk.")
• Skeletal indication of time and place often held in past time ("once there was..."); little relationship to other narrative elements	• Beginning sequence of events, but occasional out-of-sync occurrences; events without problem, problem without resolution, or little emotional response
• 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)	• Single, linear episode with clear beginning, middle, and end; the episode contains four critical elements of problem, emotional response, action, and outcome
• 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	• Plot increases in complexity with more than one episode; each episode contains problem, emotional response, action, outcome; beginning relationship between episodes
• 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.")	• 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
• 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	• Overarching problem and resolution supported by multiple, episodes; rich variety of techniques (building suspense, foreshadowing, flashbacks, denouement) to manipulate sequence

Figure 1. (cont.) Narrative rubric.



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

Figure 1. (cont.) Narrative rubric.

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. *Once 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.*

Figure 2. *Writing What You Read*: Overall effectiveness—How are features integrated in this narrative?

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 time, 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, "Turn 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.*

Figure 2. (cont.) *Writing What You Read*: Overall effectiveness—How are features integrated in this narrative?

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

Figure 2. (cont.) *Writing What You Read*: Overall effectiveness—How are features integrated in this narrative?

OUR STUDY

WWYR provided us an opportunity to examine the potential for a classroom assessment framework to serve the purposes of large-scale, rubric-based writing assessment. What technical qualities will assure the value of a rubric for large-scale writing assessment?

The overarching concern is the appropriateness of scores or assessments for their intended purposes. That concern has been historically fragmented into a melange of aspects of validity—content validity, predictive validity, concurrent validity, face validity, construct validity, criterion validity, and others—each with its own operational definition and associated methodology. We have adopted Messick's recent unified conception of validity as characterized by an expanded definition of *construct validity*. According to Messick (1992), ". . . construct validity is based on an integration of any evidence that bears on the interpretation or meaning of the test scores—including content- and criterion-related evidence" (p. 1491). Messick goes on to state that "the process of construct validation evolves from [these] multiple sources of evidence a mosaic of convergent and divergent findings supportive of score meaning (p. 1492)." Note that this conception of validity does not replace the traditional trinity of validity concepts (concept, criterion, and construct validities), but rather subsumes them and reorganizes them within a hierarchical structure.

Establishing the validity of an assessment instrument then is a process which is inherently multimethodological, with the types of evidence presented and the means for obtaining that evidence dependent on the specific nature of the domain being assessed and the purposes of the assessment. Guided by this characterization of validity, we present a mosaic of evidence for the validity of the WWYR rubric that we feel is appropriate to the purposes and intended uses of rubrics for large-scale writing assessment. Though our evidential basis is not exhaustive, the aspects we have included and the methods we have applied constitute a good core set of evidence tied to our particular purposes.

While the unification of the concept of validity under this expanded notion of construct validity provides a convenient organizational structure—highlighting the essential interrelatedness of all the various aspects of validity—when we come down to cases, it becomes desirable to decompose the unified validity construct into its subcomponents. The first subcomponent focuses attention on *content validity*, or how well the assessment samples the domain that is being assessed. Issues related to content validity are addressed largely through expert judgments of how well the assessment matches the purposes and uses of the results of the assessment. Evidence relating to the content validity of the WWYR rubric was presented earlier within the context of our discussion of its genesis—its evolution from careful study of the technical and stylistic qualities of effective narrative writing (S.A. Wolf & Gearhart, 1993a, 1993b).

The second major subcomponent is *criterion-related validity*, or how well the scores on the assessment are related to the particular construct being assessed. The major threat to this type of validity is what Messick terms "construct irrelevant variance," a concept which focuses attention on how reliably the scores on an assessment capture the essential nature of the construct without being clouded by factors extraneous to the purposes of the assessment. A potentially troubling source of construct irrelevant variance in any assessment methodology that relies on scoring rubrics applied by raters is that variance introduced by raters themselves, and we have examined the problem of rater-introduced variance in three complementary ways: through proportions of interrater agreement, through reliability indices derived from correlations between raters, and through the application of generalizability theory. As we present these results, we discuss how each of these approaches provides somewhat different information regarding the effects of raters on the variance of the test scores.

The area of criterion-related validity also subsumes questions of *convergent validity*, which is measured by the degree to which scores on an assessment correlate with other criterion measures related to the same construct, and *divergent validity*, or how capably scores on the assessment are able to discount alternative hypotheses for students' performance. We obtained evidence of convergent validity quantitatively through correlations between the WWYR rubric and an established rubric that has been used extensively in large-scale assessments, and qualitatively through our examination of raters' perceptions of the utility of the WWYR rubric. We have presented some qualitative evidence regarding the divergent validity of the WWYR rubric as a measure of narrative writing ability earlier in our section describing how the WWYR rubric was tailored to the particular features of good narrative writing. It is difficult, however, to establish divergent validity quantitatively in a study of this size. Some examples of potential alternative hypotheses would include gender bias, cultural issues, effects due to particular prompts, and others. To effectively discount these alternative hypotheses would require further studies on scales larger than the study presented here, and further work in this area is warranted. While evidence related to convergent and divergent validity is essential to establishing construct validity, such evidence by no means exhausts the set of potential sources. The case for construct validity of an assessment instrument is bolstered by any other evidence that contributes to the understanding of score meaning relative to the purposes of the assessment. Guided by this philosophy, we looked at patterns of performance across grade levels. The WWYR rubric is intended to be a developmental rubric with utility for informing instruction, and as such it should be sensitive to the development of writing competence.

Other focal areas that Messick (1992) identifies as crucial to validity but not subsumed under the umbrella of construct validity are the *relevance and utility* of the rubric for its intended purpose, the *value implications* of the use of the rubric, and the possible *social consequences* of using this assessment methodology. In the interest of establishing relevance and utility, we have elicited testimony from the raters as to the ease of use of the rubric for large-scale assessment (utility), and its capability for capturing what is important in narrative writing (relevance). Issues related to value implications focus on the possible effects of meanings of scores derived from using the assessment; here we examine evidence from raters' reflections relative to the usefulness of scores from the WWYR rubric as devices that can be used to inform instruction in writing. Finally, we consider social consequences of decisions about mastery/nonmastery based on scores derived from the rubric, by examining the stability and meaning of decisions of mastery based on different cutpoints.

METHOD

Data Sets

The narrative samples were collected from an elementary school located in a middle-class suburb. Narratives were sampled from classroom writing in Grades 1 through 6. Students' names and grade levels were removed and replaced with identification numbers. Narratives 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.

Comparison Rubric

The comparison rubric, derived from analytic scales used in the International Association for the Study of Educational Achievement Study of Written Composition (IEA) comparative studies of student writing competence, is a holistic/analytic scheme (Figure 3). In annual use in assessments of students' narratives in a California school district, this rubric has also been used extensively in our Center for evaluations of elementary students' writing (e.g., Baker, Gearhart, & Herman, 1991; Gearhart, Herman, Baker, & Whittaker, 1992; Herman, Gearhart, & Baker, 1994). Consistently demonstrating excellent levels of rater agreement and meaningful relationships with indices of instructional emphasis, the rubric represents a sound technical approach to writing assessment. Four 6-point scales are used for assessment of General Competence, Focus/Organization, Elaboration, and Mechanics; in this study, we were concerned just with narrative content, and the raters did not apply the Mechanics scale.

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

Figure 3. Comparison Narrative Rubric.

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

Figure 3. (cont.) Comparison Narrative Rubric.

Raters

Our five raters were drawn from three communities. Two raters were elementary teachers with 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. Two raters were elementary teachers experienced with other large-scale efforts; one scored elementary narrative and persuasive writing samples in English and Spanish for two years as part of a program evaluation, and the other scored writing samples of elementary school students in English and Spanish as part of a nationally implemented supplemental education program. The fifth rater was a research assistant with experience scoring elementary narrative and persuasive writing samples in English and Spanish for program evaluation.

Rating Procedures

In conducting the narrative scoring, raters were informed that the samples would represent primary (Grades 1–2), middle (Grades 3–4), or upper (Grades 5–6) elementary levels, and that sets would be labeled by levels. Raters completed comparison rubric scoring before undertaking Writing What You Read scoring. (This decision was made on the basis of both design and cost: While order of rubric is a variable that could impact judgments, we chose to focus this initial investigation on a comparison of raters' judgments with two rubrics, and we wanted the raters to focus intensively on one rubric at a time. A design that counterbalanced rubric order would have required a much larger sample of papers.)

Each phase of scoring began with (a) study and discussion of each rubric, (b) establishment of benchmark papers distributed along the scale points (based on raters' collaborative scorings of 4 to 6 unscored papers that we provided for this purpose), and (c) independent scoring of at least three papers in a row where disagreement among raters on any scale was not greater than 0.5. Raters requested and were granted permission to locate ratings at midpoints in addition to defined scale points. Training papers for each major phase were drawn from all levels. When raters began the scoring of a given level, they conducted an additional training session; raters scored preselected papers independently, resolved disagreements through discussion, and placed these "benchmark" papers in the center of the table for reference.

Because the set of papers for Grades 3 and 4 was by far the largest, raters rated half of these first, followed by Primary, Upper, and then the remaining Middle papers. Raters revisited the Middle-level benchmark papers when scoring the second half of that set. 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 entered rapidly, and a third rater rated any paper whose scores on any scale differed by more than 1 scale point. A check set of three to eight papers was included halfway through the scoring session; any disagreements were resolved through discussion which made certain that raters were not changing their criteria for scoring.

Rater Reflection

Raters were interviewed at two points in the rating process—following comparison rubric ratings (a focus group discussion) and following completion of ratings with both rubrics (an interview with pairs of raters) (see Appendix). At each interview, raters scored sample narratives and discussed the fit of the rubrics to the papers. Interviews were audiotaped and transcribed for content analysis.

RESULTS

Variance Due to Raters: Reliability of Scores

The rater contribution to score variance—a potential threat to criterion-related validity—was examined using three complementary approaches: percent agreement, correlation coefficients, and generalizability coefficients. Although the generalizability approach provides the most potentially powerful treatment of the issue of reliability, in the interest of providing as complete a mosaic of evidence as possible, we present the more traditional approaches along with caveats about the limitations on their interpretability. These analyses of agreement, correlation coefficients, and generalizability coefficients were based only on the material rated independently and thus excluded ratings negotiated during the training or the check sets.

Percentages of Agreement.

Because raters utilized midpoint ratings, percent agreement was computed for ±0, ±0.5, and ±1.0. Agreement indices were computed for each pair of raters, and those results were averaged across all the rater pairs. Agreement indices for the WWYR rubric are presented in Table 1, and indices for the comparison rubric are presented in Table 2. Rater agreement for both sets of ratings was generally satisfactory, although patterns of rater agreement differed between rubrics. While agreement for WWYR was somewhat higher and more consistent than agreement for comparison ratings, all ratings were somewhat lower than the very high rates of agreement we have obtained for the comparison rubric in prior studies (Baker et al., 1991; Gearhart et al., 1992). There were no consistent differences among rater pairs in levels of agreement, nor any evident patterns among the scales in levels of agreement.

The average percentages of agreement should be considered to be descriptive information rather than strong evidence of reliability, since, given the small range of possible values and the restricted number of scale points, rather high levels of agreement may be expected just based on chance alone. Indeed, repeated estimation of agreement indices after random permutations of the data indicated that, for these scales and these data, the chance levels of agreement for uncorrelated ratings were on the order of .16, .44, and .67 for the ±0, ±0.5, and ±1.0 indices, respectively. The introduction of very moderate correlations between ratings are sufficient to cause the percentages of adjacent (±1.0) agreement to approach the ceiling value of 1.00.

Pearson Correlations.

The average correlations (across rater pairs) for the Overall, Character, and Communication scales for the WWYR rubric (Table 1) are quite comparable to those obtained for the three scales for the comparison rubric (Table 2), while those for the Theme, Setting, and Plot scales were somewhat lower. The Setting scale was particularly problematic, with an average correlation of .48. There was less variation in correlations across rater pairs for the WWYR rubric, although this may be due largely to more stable estimates resulting from the larger number of papers that were scored using the WWYR rubric. As an example, for the comparison rubric,

TABLE 1.
Summaries of Interrater Agreement Indices for the WWYR Rubric: Means and Standard Deviations of Agreement Indices Across the Ten Pairs of Raters.

	Overall Effectiveness	Theme	Character	Setting	Plot	Communication
Pearson Correlations						
M	.64	.59	.66	.48	.57	.66
SD	.10	.10	.12	.14	.10	.10
% Agreement ± 0						
M	.46	.41	.43	.45	.39	.44
SD	.06	.07	.09	.10	.08	.07
% Agreement ± 0.5						
M	.85	.72	.72	.71	.76	.82
SD	.05	.06	.10	.08	.11	.06
% Agreement ± 1.0						
M	.96	.95	.95	.93	.95	.97
SD	.03	.03	.04	.04	.02	.02

Note. Sample sizes ranged from 27 (Raters 1 and 4) up to 93 (Raters 3 and 5).

TABLE 2.
Summaries of Interrater Agreement Indices for the Comparison Rubric: Means and Standard Deviations of Agreement Indices Across the Ten Pairs of Raters.

	General Competence	Focus/ Organization	Development/ Elaboration
Pearson Correlations			
M	.68	.60	.63
SD	.19	.17	.15
% Agreement ± 0			
M	.37	.28	.31
SD	.10	.14	.11
% Agreement ± 0.5			
M	.73	.64	.67
SD	.10	.12	.12
% Agreement ± 1.0			
M	.92	.92	.94
SD	.11	.08	.07

Note. Sample sizes ranged from 12 (Raters 1 and 5) up to 21 (Raters 1 and 3).

the lowest correlations were obtained for the one and five pairing of raters (.28 and .25 for the General Competence and the Focus/Organization scales, respectively); those estimates, however, were based on a sample of only 12 papers.

The average correlations can be interpreted much like classical reliability coefficients, with the difference that instead of estimating the correlation between parallel forms of a test (as in classical reliability theory), we are estimating the correlation between parallel ratings of a single test. There are two main weaknesses to this approach to estimating reliability. First, while there are no hard and fast guidelines about what constitutes an adequate level of rater agreement, in this case, most experts would probably agree that the correlations for some of the scales are somewhat smaller than might be deemed acceptable. Unlike generalizability theory (following), this approach to estimating reliability provides no recourse or prescription for improving that situation. Second, since correlations provide information about the relative rankings of individuals, it is possible to have a high correlation without necessarily having good agreement between raters. Raters might agree very well on the *relative* ranking of individuals without agreeing on where those individuals stand compared to some *absolute* standard for performance. This issue becomes crucial in situations in which we value comparability of absolute scores across different raters, as is typically the case for large-scale assessments.

Generalizability Coefficients.

Generalizability theory is a powerful methodology for addressing issues of rater agreement (Brennan, 1984; Crocker & Algina, 1986; Shavelson & Webb, 1991), and it has potential for addressing the deficiencies alluded to earlier regarding simple correlations. Generalizability theory is much more flexible than classical reliability theory in that generalizability coefficients can be tailored to suit the particular purposes of an evaluation. For example, separate generalizability coefficients can be computed for relative and absolute decisions: If one is interested mainly in accurately ranking a set of essays, then a relative coefficient would be of interest; on the other hand, if one is making decisions about proficiency by comparing scores to an absolute standard, such as a cut score, or is comparing scores assigned by different raters, then an absolute coefficient is more appropriate.

Generalizability coefficients are ratios of the variance due to the objects of measurement (in our case, students' essay scores) to the total variance due to the objects of measurement and the conditions of measurement (in our case, raters). Table 3 shows the proportions of variance attributable to essays, raters, and to the essay by rater interaction, and the resultant generalizability coefficients. Coefficients for both relative and absolute decisions are reported. Note that for both rubrics the proportion of variance due to raters is almost negligible. This indicates quite good consistency in the application of the scoring rubrics across raters and has very positive implications

TABLE 3.
Generalizability Coefficients

		Variance Components			Generalizability Coefficients	
Rubric	**Scale**	**E**	**R**	**ER**	**Relative**	**Absolute**
	General Competence	.68	.00	.32	.68	.68
Comparison	Focus/ Organization	.63	.01	.36	.64	.63
	Development/ Elaboration	.66	.01	.34	.66	.65
	Overall	.60	.01	.40	.60	.59
	Theme	.55	.04	.41	.57	.55
WWYR	Character	.62	.01	.37	.63	.62
	Setting	.47	.00	.53	.47	.47
	Plot	.55	.00	.45	.55	.55
	Communication	.62	.00	.37	.63	.63

Note. The table contains standardized variance component estimates for essay (E), rater (R), and the essay by rater interaction (ER), and the generalizability coefficients derived from those estimates, for each of the comparison and WWYR scales.

with respect to the feasibility of using scores based on these rubrics to make absolute decisions about students' proficiencies, such as assignments to proficiency categories based on cutpoints, or comparisons of scores assigned to students by different raters. If the variance due to raters were large, then we would have very little assurance that scores assigned to students by different raters indicated true differences in competence or instead reflected differences in raters' interpretation and anchoring of rubric scale points. This is not the case here, however, and the very small variance components for raters ensure that the generalizability coefficients for relative and absolute decisions will be quite close together, as we see in Table 3.

Comparisons Across Rubrics.

Comparing across rubrics and scales, we see that the G-coefficients for the comparison rubric scales tend to be consistently higher than those for the WWYR rubric. G-coefficients for the comparison rubric are quite consistent across scales, while there is considerable variation in the generalizability for the WWYR scales, with the Setting scale the most problematic with an estimated generalizability coefficient of .47.

D-Study Coefficients.

The results of a generalizability study (G-study) can be extended to what is called a decision study (D-study). In classical test theory, the reliability of the test is a function of the length of the test; longer tests are more reliable, and the reliability of a test can be improved by adding more items. The analogous procedure in a rating situation is to improve reliability by adding more raters, multiply scoring each essay, and aggregating the results. The G-study coefficients from Table 3 can be interpreted as reliability indices for scores based on a single rater. If those coefficients are too low, then a D-study can be done to examine the effects on generalizability of adding more raters. An informed decision can then be made as to how many raters should be used to attain adequate levels of generalizability.

If we compare the results in Table 3 with those in Tables 1 and 2, we see that the generalizability coefficients agree closely with the average Pearson correlations. Although there are no cut-and-dried guidelines for what determines an adequate level of reliability, most researchers would probably like to see reliabilities of at least .75, and the generalizability coefficients for both rubrics fall well below that threshold. The next step within the context of generalizability theory was to use the results of the G-study to perform a D-study in order to determine how to attain an acceptable reliability level. Table 4 reports D-study generalizability coefficients for scores based on 1, 2, 3, and 5 raters.

The results of the D-study show that for all of the comparison scales and for three of the WWYR scales, adequate reliability (as defined previously)

TABLE 4.
D-Study Coefficients

		Relative				Absolute			
Rubric	Scale	1	2	3	5	1	2	3	5
Comparison	General Competence	.68	.81	.86	.91	.68	.81	.86	.91
	Focus/ Organization	.64	.78	.84	.90	.63	.77	.84	.89
	Development/ Elaboration	.66	.80	.85	.91	.65	.79	.85	.90
WWYR	Overall	.60	.75	.82	.88	.59	.75	.81	.88
	Theme	.57	.73	.80	.87	.55	.71	.79	.86
	Character	.63	.77	.83	.89	.62	.77	.83	.89
	Setting	.47	.64	.73	.82	.47	.64	.73	.82
	Plot	.55	.71	.79	.86	.55	.71	.79	.86
	Communication	.63	.77	.83	.89	.63	.77	.83	.89

Note. D-study generalizability coefficients for relative and absolute decisions for essay scores based on 1, 2, 3, or 5 raters.

can be obtained through the use of two raters. Note, however, that for the WWYR Setting scale, even the use of three raters is not sufficient to ensure a reliability level of .75. Using four raters would result in a coefficient of .78 for this scale. Again, due to the very small proportions of variance attributable to the rater main effects, results and interpretations for relative and absolute decisions are nearly identical.

Other Evidence for Criterion-Related Validity

This section contains four analyses of the Writing What You Read rubric's capacity to produce meaningful results: (a) comparisons of students' scores across grade levels (scores should increase with grade level); (b) intercorrelations of scales within rubrics (for each rubric, scales should not be highly correlated); (c) correlations of ratings across rubrics (WWYR scores should correlate significantly with comparison scores); (d) an analysis of decision consistency across rubrics (raters should make similar decisions about students' competence across rubrics). All ratings contributed to these results: Paper scores were computed as the average of the independent ratings *or* the resolved score achieved through discussion during the training and check sets.

Grade-Level Comparisons.

These comparisons are intended to examine what might be considered the "developmental validity" of the two rubrics: If we accept the premise that students' writing proficiency increases with age, then a rubric that captures

TABLE 5.
Descriptives, Comparison Rubric

	Scale		
Level	General Competence	Focus/ Organization	Development/ Elaboration
Primary (n = 16)			
M	2.05	2.29	2.27
SD	0.47	0.48	0.45
Middle (n = 36)			
M	2.58	2.68	2.79
SD	0.55	0.50	0.59
Upper (n = 17)			
M	3.54	3.66	3.67
SD	0.49	0.67	0.57

Note. For this analysis, n = number of subjects. ANOVAs examined differences among levels for each scale: General Competence, $F(2, 66) = 36.38, p < .0001$; Focus/Organization, $F(2, 66) = 29.14, p < .0001$; Development/Elaboration, $F(2, 66) = 26.98, p < .0001$.

this proficiency should show scores that increase with grade level. Tables 5 and 6 contain descriptive statistics for each rubric and, for each scale, the results of ANOVAs by level. For each rubric, there were score differences in the expected direction by grade level. The pattern of score differences was the same for all scales and both rubrics, although the ANOVA result for one WWYR scale (Plot) was not significant.

TABLE 6.
Descriptives, *Writing What You Read* Rubric

	Scale					
Level	Overall	Theme	Character	Setting	Plot	Communication
Primary (n = 17)						
M	2.29	2.47	2.15	2.27	2.44	2.33
SD	0.39	0.48	0.53	0.42	0.49	0.44
Middle (n = 36)						
M	2.50	2.61	2.40	2.49	2.55	2.51
SD	0.44	0.45	0.53	0.43	0.47	0.49
Upper (n = 20)						
M	2.87	3.02	2.78	2.73	2.80	2.96
SD	0.59	0.64	0.74	0.51	0.64	0.64

Note. For this analysis, n = number of subjects. ANOVAs examined differences among levels for each scale: Overall, $F(2, 70) = 7.11, p < .002$; Theme, $F(2, 70) = 6.11, p < .004$; Character, $F(2, 70) = 5.45, p < .006$; Setting, $F(2, 70) = 4.93, p < .01$; Plot, $F(2, 70) = 2.47, p < .092$; Communication, $F(2, 70) = 7.52, p < .001$

Intercorrelations of Scales Within Rubrics.

Tables 7 and 8 contain intercorrelations of scales for each rubric. All scales were highly correlated, indicating that raters were not making highly differentiated judgments about a narrative's competence along each dimension. Based on these results, scales for both rubrics are not empirically distinct.

Correlations of Ratings Across Rubrics.

Table 9 contains intercorrelations of scales across rubrics. Across rubrics, scores were highly intercorrelated, although the correlations were lower in magnitude than the within-rubric correlations (Tables 7 and 8).

Decision Consistency Across Rubrics.

To examine consistency in raters' judgments of narrative competence across rubrics, we cross-classified scores for General Competence (comparison) and Overall Effectiveness (WWYR) (Table 10). These results must be interpreted in the context of two important issues. First, although both rubrics are 6-point scales, their scale points do not correspond in meaning; in particular, the WWYR rubric is developmental and is not intended to locate competency at any particular level. Second, although the "best fit" for WWYR's definition of a competent narrative may be Level 3 ("One episode narrative (either brief or more extended) which includes the four critical elements of problem, emotional response, action, and outcome. . . ."), the criteria for this level were considered unclear by our raters (discussion follows).

TABLE 7.
Scale Correlations, Comparison Rubric (N = 184)

	Scale		
Level and Scale	**General Competence**	**Focus/ Organization**	**Development/ Elaboration**
Primary (n = 36)			
General Competence		.80*	.81*
Focus/Organization			.74*
Middle (n = 115)			
General Competence		.87*	.90*
Focus/Organization			.80*
Upper n = 35)			
General Competence		.91*	.86*
Focus/Organization			.82*
Overall (n = 184)			
General Competence		.91*	.92*
Focus/Organization			.85*

*p < .001.

TABLE 8.
Scale Correlations, *Writing What You Read* Rubric ($N = 187$)

	Scale					
Scale	Overall	Theme	Character	Setting	Plot	Communication
Primary (n = 37)						
Overall		.88*	.86*	.86*	.86*	.86*
Theme			.85*	.73*	.87*	.83*
Character				.77*	.82*	.81*
Setting					.77*	.82*
Plot						.82*
Middle (n = 112)						
Overall		.92*	.91*	.87*	.93*	.94*
Theme			.88*	.81*	.89*	.89*
Character				.85*	.88*	.88*
Setting					.82*	.81*
Plot						.92*
Upper n = 38)						
Overall		.94*	.90*	.92*	.95*	.97*
Theme			.90*	.91*	.93*	.95*
Character				.83*	.91*	.91*
Setting					.89*	.92*
Plot						.94*
Total (n = 187)						
Overall		.93*	.91*	.89*	.93*	.94*
Theme			.90*	.84*	.90*	.90*
Character				.84*	.89*	.89*
Setting					.84*	.85*
Plot						.91*

*$p < .001$.

We chose a WWYR mean rating of 3.0 or above as evidence of competence, and compared WWYR judgments against comparison rubric ratings of 3.5 or above, consistent with the comparison rubric's distinction between a "developing writer" (Level 3) and a "competent writer" (Level 4). Most papers were judged as lacking in competence. Raters agreed in their classifications of 146 of 176 papers (Pearson $\chi^2 = 46.69, p < .0001$). However, there was no consistent agreement in classification of "competent" papers: Of the 55 papers judged as competent with either rubric, only 25 were classified as competent with both rubrics.

The results of the decision consistency analysis serve two purposes. First, to the degree that raters' decisions are consistent across rubrics, we have evidence for the convergent validity of the WWYR rubric relative to the alternative measure provided by the comparison rubric. But note also that we have evidence of divergent validity in the substantial disagreement on what constitutes a proficient writer. This aspect is important when we

TABLE 9.
Correlations Across Rubrics

Comparison Scale	WWYR Scale					
	Overall	Theme	Character	Setting	Plot	Communication
Primary (n = 36)						
General Competence	.62***	.61***	.70***	.59***	.69***	.68***
Focus/Organization	.46*	.54**	.44*	.43*	.56***	.58***
Development/ Elaboration	.62***	.60***	.61***	.65***	.65***	.64***
Middle (n = 107)						
General Competence	.79***	.75***	.75***	.71***	.72***	.77***
Focus/Organization	.71***	.68***	.65***	.60***	.66***	.68***
Development/ Elaboration	.74***	.71***	.70***	.65***	.70***	.74***
Upper (n = 33)						
General Competence	.74***	.71***	.72***	.68***	.74***	.73***
Focus/Organization	.65***	.55***	.59***	.56***	.65***	.64***
Development/ Elaboration	.67***	.62***	.71***	.60***	.65***	.68***
Total (n = 176)						
General Competence	.75**	.73**	.74**	.66**	.67**	.74**
Focus/Organization	.67**	.66**	.64**	.58**	.62**	.68**
Development/ Elaboration	.72**	.70**	.71**	.64**	.66**	.73**

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

consider the possible social consequences of implementing a new scoring system. Individuals who would be qualified under the old system might be judged underqualified by the new system, a situation likely to engender dissension and controversy.

TABLE 10.
Cross-Classification of Comparison and WWYR Scores (N = 176)

WWYR Overall Effectiveness	Comparison General Competence	
	< 3.0	= or > 3.0
< 2.5	121	14
= or > 2.5	16	25

Note. For each rubric, each paper was scored by at least two raters; paper scores were computed as the mean of all raters' judgments.

Raters' Reflections

Raters raised issues regarding the *relevance, utility*, and *value* of the WWYR rubric.

Relevance: Representation of Narrative Content.

Raters examined carefully the ways that each rubric did or did not capture important qualities of narrative writing. Overall, WWYR was viewed as more comprehensive in its analysis of narrative as well as more "positive" in each of its scale-point definitions—more specific about narrative qualities that a piece does contain, less "negative" regarding what a piece does not contain. Raters also welcomed what they perceived as WWYR's more complete analysis of a narrative's "development." Indeed, they reported adding to the comparison Development/Elaboration scale content which they considered central to a judgment of narrative. One rater explained, "I put feeling under Elaboration. I know it's not, but . . . you need to." Another rater commented,

> There's a big difference between actually seeing something visually [the emphasis of the comparison rubric] and feeling something. . . . Something can be "vivid," and something can be "elaborate," but it might not make you feel emotionally.

In their critique of WWYR content, raters focused on Plot, Overall Effectiveness, Communication, and the absence of a scale like the comparison rubric's Focus/Organization scale. Plot and Overall Effectiveness were seen as weak at levels two through four, handling ineffectively those longer narratives that contained a series of incomplete episodes. Communication was considered helpful in pinpointing particular techniques, but its emphasis on language choices "appropriate to the narrative" made it difficult for the raters to give a child credit for stylistic strength that did not necessarily contribute to the narrative. In addition, they felt that Communication could be differentiated—at least for instructional applications—as separate scales for style, tone, and voice. (An early version of WWYR contained these dimensions. See S.A. Wolf & Gearhart, 1993a, 1993b, for explication of these components.) Finally, raters felt that WWYR needed a scale in some way analogous to the comparison rubric's Focus/Organization scale. While seen as rather dry and perhaps expositionlike, this scale captured for these raters a dimension of organizational competence missing in WWYR.

Raters felt that neither rubric was able to capture a narrative's local strengths: "Maybe they have one character description, or a setting, or something funny, and you laugh, but it really doesn't allow itself to be 4 and you want to tell them, 'Hey, you made me laugh here, or look at all these similes you were using.' " Similarly, some raters felt that neither rubric represented creativity very well: "There might be some idiosyncratic quality or some uniqueness about it, some originality that you can't really score." Wanting to "give credit" to a child for a moment of insight, humor, language use, or cleverness, they suggested providing a place on the rating form for personal comments to each writer on strengths and weaknesses.

Utility: Ease of Use and Feasibility of Use for Large-Scale Assessment. Although most raters felt that application of the WWYR rubric was a slower, more "analytical" process than comparison rubric rating, only one of the five raters remained uncomfortable: "[The WWYR rubric is] so broken apart, analytic, that it confuses me." Indeed, the WWYR rubric did contain a greater number of scales and detail at each scale point, and, for this rater, the constructs required explication ("explicit and implicit, didactic and revealing . . . it's too much to keep track of"). For the remaining raters, the acknowledged difficulty of WWYR scoring was balanced with enjoyment "because [WWYR] talked about the different subtleties of language and the different styles and emotions that you could use to make it more sophisticated and improve it. Whereas the comparison [rubric] didn't really give that feeling . . . language . . . just seemed like a skill rather than a quality of the work."

Raters also appreciated the specificity of the WWYR rubric. Four of the five raters reported difficulty anchoring their comparison rubric judgments based on relative criteria: "This 'few, many, little, and more' kind of vocabulary . . . was really a problem in the beginning . . . What is 'many?' What is 'few?' We had to make our own kind of interpretations, and then compare as we went on reading." Wishing for more positive and specific descriptions, one rater commented: "What is the paper *do*ing, even though there might be inappropriate [language]. . . . 'No development of narrative elements'—what can you say instead of that?" To adapt, raters reported several strategies for resolving uncertainty: expanding the list of comparison rubric criteria (e.g., the addition of "emotion" to Development, as discussed earlier); making iterative comparisons with higher and lower scale points; using the anchor terms (e.g., 1: Minimal Evidence of Achievement/Insufficient Writer; 3: Some Evidence of Achievement/Developing Writer, etc.); making an initial dichotomous judgment between "Developing" (1–3) and "Competent" (4–6) writer and then refining the decision. The raters felt that WWYR, in contrast, supported greater focus on the fit of a narrative to the characteristics listed at a given level.

Raters agreed that the comparison rubric had the capacity to be used reliably and with reasonable speed for large-scale assessment purposes. In contrast, the feasibility and utility of WWYR for large-scale assessment were left as unanswered questions. First, although raters acknowledged that they themselves had acquired expertise with WWYR in half a day, they nevertheless expressed concern about the rater training that would be required to implement a large-scale program based on WWYR assessment. Second, although the raters considered the WWYR Overall Effectiveness scale as a possible holistic replacement for comparison's General Competence, they were concerned about the relation between the two judgments: Overall Effectiveness required a rater to judge the narrative's integration of other narrative elements, still a fairly analytic task that felt different in content and in process from a General Competence decision. Raters suggested improvements of the WWYR rubric that they felt would have facilitated scoring for them: highlighting key terms, listing criteria as bullets, and adopting overarching descriptors like those in the comparison rubric's left column (e.g., Developing Writer, Competent Writer).

Value: Instructional Potential.

Most raters viewed the WWYR rubric as having far more instructional potential than the comparison rubric, and those four raters who were classroom teachers planned to utilize it in some form in their classrooms.

> [WWYR] allows you to compliment other strengths, and their styles. . . . It's wonderful to have it for a teacher resource to direct the children, and the parents. . . . When I'm scoring kids [with the comparison rubric], I'm having a hard time putting into words what I want them to do. With WWYR, I could get up and directly teach a lesson.

But one of the four teachers felt that WWYR demanded more analysis than she could routinely or profitably undertake in the classroom. For this rater, difficulty of use limited instructional potential: "For many teachers, you have to give them something that's easy to apply, an easy tool that we can use. . . . Not too much analyzing, not too much re-reading. Something automatic. I would like a tool like that . . . for our daily writing." A rubric with content as complex as WWYR might be useful, she granted, when undertaking "a major project, then I want to use something like the *Writing What You Read*, if I want to touch on every single part [of the writing]."

SUMMARY AND DISCUSSION

Is there a way to bridge the gap between writing assessments that support effective instruction and those that are useful for large-scale assessment? We have addressed this question by examining the potential for a classroom assessment framework to serve the purposes of large-scale writing assessment. The design of the *Writing What You Read* (WWYR) narrative rubric began in the classroom, prompted by the need for assessments that "chart . . . the course between uniformity of judgment on the one hand and representation of complexity and diversity on the other hand" (D.P. Wolf, Bixby, Glenn, & Gardner, 1991). Existing narrative rubrics designed for large-scale assessment did not, in our view, have this capacity. Consider Grant Wiggins' (1993) example of a rubric that gives a story the highest score if it "describes a sequence of episodes in which almost all story elements are well developed (i.e., setting, episodes, characters' goals, or problems to be solved). The resolution of the goals or problems at the end are [sic] elaborated. The events are represented and elaborated in a cohesive way." Wiggins comments, "Surely this is not the best description possible of a good story (p. 201)." Surely not. But could the "best description," or even a better description, be captured in a technically sound rubric?

WWYR was designed as an alternative to narrative rubrics that are not grounded in *genre*, either in its traditional sense of a classification system for organizing literature (a system much subject to change) or in its more current sense of social action constrained by particular rhetorical forms. The rubric contains five analytic scales for Theme, Character, Setting, Plot, and Communication, and a sixth, holistic scale for Overall Effectiveness, and each scale contains six levels designed to match current understandings of children's narrative development. The development of character, the symbolism in setting, the complexity of plot, the subtlety of theme, the selected point of view, and the elaborate use of language all depend on and are defined by genre. Thus, the WWYR rubric was designed as a framework to support the coordination of assessment with instruction: If we are going to teach children about narrative and how to grow as young story writers, then surely we want to use more precise language and describe a fuller picture of what narrative is, to provide them access to more intriguing and more authentic possibilities. The WWYR rubric is a simplification, yet its language and focus provide a key to a much larger door, opening onto the evocative, emotional, and eminently human symbol system of narrative meaning.

To guide the development of a model for examining the WWYR rubric's technical quality, we adopted Messick's recent unified conception of validity (Messick, 1992). Under this framework, traditional constructs of *content validity* and *criterion-related validity* are subsumed by an expanded definition of *construct validity*, and attention is also focused on the relevance and utility, value implications, and social consequences of the assessment. The overarching concern here is the appropriateness of scores or assessments for their intended purposes, and, therefore, the methods for establishing the validity of an assessment instrument vary depending on the specific nature of the domain being assessed and the purposes of the assessment. Our study addressed many of the components that Messick argues are important evidence.

Evidence for content validity was provided by previously published analyses of the WWYR rubric's content (S.A. Wolf & Gearhart, 1993a, 1993b), including its relation to existing frameworks for understanding narrative writing and its development. In examining criterion-related validity, we looked first at the variance introduced by raters, a potential threat to criterion-related validity. We found that while neither the WWYR or comparison rubrics met commonly accepted standards of reliability for scores based on a single rater, acceptable levels of reliability for most scales could be achieved by doubly rating essays. Other evidence of criterion-related validity was provided by findings regarding patterns of (a) scores across grade level and (b) convergent and divergent validity. First, the scores from both rubrics produced a pattern of increasing competence with grade level. Second, WWYR scores were highly correlated with the comparison scores; evidence for the distinctiveness of the two scales was provided by the finding that cross-rubric scale correlations were lower than within-rubric scale correlations.

Evidence for the relevance, utility, and value of the WWYR rubric was obtained from the raters' reflections on their experience as judges. The raters felt that the content of WWYR captured more aspects of narrative than the comparison rubric, although they recommended revisions of the scales for Plot, Communication, and Overall Effectiveness, as well as the addition of a scale like the comparison rubric's Focus/Organization scale. Regarding the utility of WWYR, the raters expressed concern about the professional development that would be required for scoring in the large-scale context, despite their recognition that they had achieved understandings of WWYR and reasonable consensus in its use after only a half-day training session. The raters were agreed that WWYR had considerably greater instructional potential than the comparison rubric, and they planned to utilize WWYR in their own classrooms.

Evidence relevant to the potential social consequences of WWYR use was provided by analyses of the decisions raters made regarding students' competence. Comparisons of raters' judgments made with both rubrics for the same narratives indicated some consistency in their decisions, although disagreements in classifications of "competent" narratives suggested distinctive definitions for competence. These results underscore the need for research on the potential impact of new assessments on high-stakes decisions. If a school district, for example, is considering adoption of a new measure—whether WWYR or any other—it is critical that they examine how cutpoints on either measure influence decisions regarding a student's mastery. A change in assessments could otherwise have serious social consequences.

Thus, our technical study has produced evidence that at least three scales of the *Writing What You Read* narrative rubric—an analytic writing rubric designed to enhance teachers' understandings of narrative and to inform instruction—can be used reliably and meaningfully in large-scale assessment of elementary-level writing, provided that each narrative is rated by two raters. While we would have preferred that our analyses yield evidence of the technical soundness of all six scales, it is nevertheless heartening that any scales as substantive as WWYR's could produce findings this positive in an initial study. However, consistent with other studies of analytic scales, neither the WWYR nor the comparison rubric produced patterns of highly distinctive scale judgments. While 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, our quantitative findings suggest that WWYR scale judgments may not provide a technically sound profile of students' strengths and weaknesses.

We do not view these findings as a basis for rejecting an analytic *framework* for scoring. Further research is needed to determine the factors that support or constrain distinctive scale judgments—the structure and content of analytic rubrics, the types of material to be rated, and the methods of rater training. If technical studies continue to demonstrate that scale judgments cannot be distinguished from overall competence ratings, we would maintain that validity issues regarding the relevance, value, and social consequences of assessments constitute arguments for designing "analytic" alternatives to holistic scoring. One option might be assignment of a single score supplemented with rater commentary on strengths and weaknesses, commentary that could be guided by "analytic" prompts or checklists. In this context, it is heartening to note that Huot (1993) found that holistic rubrics can serve as frameworks that support a wide variety of rater comments.

The design and evaluation of a rubric to serve the dual purposes of classroom and large-scale assessment have confronted us with the "test-maker's dilemma" (Wiggins, 1993): Rubrics capturing the complexity of accomplished writing performance in the classroom may not support technically sound assessment in the large-scale context. The challenge is to optimize the purposes of assessment at all levels in a coordinated system. We have shown that WWYR—a rubric designed to capture valued qualities of distinctive writing genres—can support both enhanced opportunities to learn in the classroom as well as the validity of ratings in the large-scale context. This article conveys simultaneously both the results of our technical study and a model for examining the technical quality of large-scale writing assessments.

REFERENCES

Baker, E.L. (1987, September). *Time to write: Report of the US-IEA Study of Written Composition*. Invited presentation to the IEA General Assembly, Teachers College, New York.

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

Baxter, G.P., Glaser, R., & Raghavan, K. (1994). *Analysis of cognitive demand in selected alternative science assessments* (CSE Tech. Rep. No. 382). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

Brennan, R.L. (1984). *Elements of generalizability theory.* Iowa City: ACT Publications.

Broad, R.L. (1994). "Portfolio scoring": A contradiction in terms. In L. Black, D.A. Daiker, J. Sommers, & G. Stygall (Eds), *New directions in portfolio assessment* (pp. 263–276). Portsmouth, NH: Boynton/Cook.

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

Condon, W., & Hamp-Lyons, L. (1994). Maintaining a portfolio-based writing assessment: Research that informs program development. In L. Black, D.A. Daiker, J. Sommers, & G. Stygall (Eds), *New directions in portfolio assessment* (pp. 277–285). Portsmouth, NH: Boynton/Cook.

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

Elbow, P. (1994). Will the virtues of portfolios blind us to their potential dangers? In L. Black, D.A. Daiker, J. Sommers, & G. Stygall (Eds), *New directions in portfolio assessment* (pp. 40–55). Portsmouth, NH: Boynton/Cook.

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

Gearhart, M., & Herman, J.L. (1995, Winter). Portfolio assessment: *Whose work is it?* Issues in the use of classroom assignments for accountability. *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., & 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. No. 377). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

Haswell, R., & Wyche-Smith, S. (1994). Adventuring into writing assessment. *College Composition and Communication, 45*, 220–236.

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

Herman, J.L., Gearhart, M., & Baker, E.L. (1994). Assessing writing portfolios: Issues in the validity and meaning of scores. *Educational Assessment, 1*, 201–224.

Herman, J.L., & Golan, S. (1991). *Effects of standardized testing on teachers and learning: Another look* (CSE Tech. Rep. No. 334). Los Angeles: University of California, Center for the Study of Evaluation.

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

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

Huot, B. (1990a). The literature of direct writing assessment: Major concerns and prevailing trends. *Review of Educational Research, 60*, 237–263.

Huot, B. (1990b). Reliability, validity, and holistic scoring: What we know and what we need to know. *College Composition and Communication, 41*, 201–213.

Huot, B. (1993). The influence of holistic scoring procedures on reading and rating student essays. In M.W. Williamson & B.A. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (pp. 206–236). Cresskill, NJ: Hampton Press.

Huot, B. (1994). Beyond the classroom: Using portfolios to assess writing. In L. Black, D.A. Daiker, J. Sommers, & G. Stygall (Eds), *New directions in portfolio assessment* (pp. 325–333). Portsmouth, NH: Boynton/Cook.

Johnson, E.G., & Carlson, J.E. (1994). *The NAEP 1992 Technical Report.* Washington, DC: National Center for Educational Statistics.

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

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

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

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

Moss, P.A. (1994). Validity in high stakes writing assessment. *Assessing Writing, 1*, 109–128.

Murphy, S.A. (1994). Portfolios and curriculum reform: Patterns in practice. *Assessing Writing, 1*, 175–206.

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

O'Neil, J. (1993). On the New Standards Project: A conversation with Lauren Resnick and Warren Simmons. *Educational Leadership, 50*(5), 17–21.

Paul, R.W. (1993). *Pseudo critical thinking in the educational establishment: A case study in educational malpractice.* Sonoma, CA: Sonoma State University, Center for Critical Thinking and Moral Critique.

Resnick, L., Resnick, D., & DeStefano, L. (1993). *Cross-scorer and cross-method comparability and distribution of judgments of student math, reading, and writing performance: Results from the New Standards Project Big Sky scoring conference* (CSE Tech. Rep. No. 368). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

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

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

Shepard, L.A. (1991). Will national tests improve student learning? *Phi Delta Kappan, 73*, 232–238.

Simmons, W., & Resnick, L. (1993). Assessment as the catalyst of school reform. *Educational Leadership, 50*(5), 11–16.

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

Spalding, E. (1995). The New Standards Project and English Language Arts Portfolio: A Report on process and progress. *The Clearing House, 68*(4), 219–224.

Vermont Department of Education. (1991). *"This is my best": Vermont's writing assessment program, pilot year 1990–1991.* Montpelier, VT: Author.

White, E.M. (1994). Portfolios as an assessment concept. In L. Black, D.A. Daiker, J. Sommers,

& G. Stygall (Eds), *New directions in portfolio assessment* (pp. 25–39). Portsmouth, NH: Boynton/Cook.

Wiggins, G. (1993). Assessment: Authenticity, context, and validity. *Phi Delta Kappan, 75*(3), 200–208, 210–214.

Wiggins, G. (1994). The constant danger of sacrificing validity to reliability: Making writing assessment serve writers. *Assessing Writing, 1*, 129–141.

Williamson, M. (1994). The worship of efficiency: Untangling theoretical and practical considerations in writing assessment. *Assessing Writing, 1*, 147–174.

Wolf, D.P. (1993). Assessment as an episode of learning. In R. Bennett & W. Ward (Eds.), *Construction versus choice in cognitive measurement* (pp. 213–240). Hillsdale, NJ: Erlbaum.

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

Wolf, S.A., & Gearhart, M. (1993a). *Writing What You Read: Assessment as a learning event* (CSE Tech. Rep. No. 358). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

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

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

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

APPENDIX

Interview Questions

Comparison

Please rate the attached narratives on the rating sheet, and jot notes on the questions below in preparation for the interview.

Narrative title ________________________

____ General Competence
____ Focus/Organization
____ Development/Elaboration

What does this rubric capture about this narrative?
What does it not capture?
What does the General Competence scale capture about this narrative?
What does the General Competence scale not capture?
What does the Focus/Organization scale capture about this narrative?
What does the Focus/Organization scale not capture?
What does the Elaboration scale capture about this narrative?
What does the Elaboration scale not capture?

[Repeated for two additional narratives.]

WWYR

Please rate the attached narratives on the rating sheet, and jot notes on the questions below in preparation for the interview.

Narrative title ____________________

____ Overall Effectiveness
____ Theme
____ Character
____ Setting
____ Plot
____ Communication

What does the WWYR rubric capture about this narrative?
What does it not capture?
What makes WWYR 'rater friendly'—easy to apply:
What makes WWYR 'rater unfriendly'—difficult to apply:
What does the Overall Effectiveness scale capture about this narrative?
What does the Overall Effectiveness scale not capture?
What does the Theme scale capture about this narrative?
What does the Theme scale not capture?
What does the Character scale capture about this narrative?
What does the Character scale not capture?
What does the Setting scale capture about this narrative?
What does the Setting scale not capture?
What does the Plot scale capture about this narrative?
What does the Plot scale not capture?
What does the Communication scale capture about this narrative?
What does the Communication scale not capture?

[Repeated for two additional narratives.]

Compare/contrast

What are the strengths and weaknesses of each rubric for large-scale assessment?
What are the strengths and weaknesses of each rubric for classroom assessment?