Putting Rubrics to the Test: The Effect of a Model, Criteria Generation, and Rubric-Referenced Self-Assessment on Elementary School Students' Writing

Heidi L. Andrade, Ying Du, and Xiaolei Wang, *University at Albany, SUNY*

The purpose of this study was to investigate the effect of reading a model written assignment, generating a list of criteria for the assignment, and self-assessing according to a rubric, as well as gender, time spent writing, prior rubric use, and previous achievement on elementary school students' scores for a written assignment (N = 116). Participants were in grades 3 and 4. The treatment involved using a model paper to scaffold the process of generating a list of criteria for an effective story or essay, receiving a written rubric, and using the rubric to self-assess first drafts. The comparison condition involved generating a list of criteria for an effective story or essay, and reviewing first drafts. Findings include a main effect of treatment and of previous achievement on total writing scores, as well as main effects on scores for the individual criteria on the rubric. The results suggest that using a model to generate criteria for an assignment and using a rubric for self-assessment can help elementary school students produce more effective writing.

Keywords: self-assessment, rubric, formative assessment, writing

Rubrics have become popular with teachers as a means of communicating expectations for an assignment, providing focused feedback on works in progress, and grading final products (Andrade, 2000; Moskal & Leydens, 2000; Popham, 1997; Quinlan, 2006). Although educators tend to define the word "rubric" in slightly different ways, a commonly accepted definition is a document that articulates the expectations for an assignment by listing the criteria, or what counts, and describing levels of quality from excellent to poor (Andrade, 2000).

Rubrics are often used to grade student work but many authors argue that they can serve another, more important, role as well: Rubrics can teach as well as evaluate (Arter & McTighe, 2001; Quinlan, 2006; Stiggins, 2001; Stix, 1996). Stiggins argues that, when used as part of a formative, student-centered approach to assessment, rubrics have the potential to help students develop understanding and skill, as well as make dependable judgments about the quality of their own work. Identifying students as "the key assessment users" (p. 17), Stiggins notes that they should be able to use assessments in many of the same ways that teachers use them—to clarify the standards for a quality performance, and to guide ongoing feedback about progress toward those standards. Other assessment theorists, including Black and Wiliam (1998), Brookhart (2003), Shepard (2000), and Wiggins (1998) put forward a similar conception of assessment as a moment of learning.

Research on the Effects of Rubric Use

The claim that rubrics can promote learning and achievement has intuitive appeal but there is only limited empirical evidence to support it. We found two relevant quasi-experimental studies. In a study of the relationship between having access to a rubric and seventh and eighth grade students' writing (Andrade, 2001), having a rubric was associated with higher scores on only one of three essays written by the students. However, questionnaires administered at the end of the study revealed that students who had received rubrics tended to identify more of the criteria by which their writing was evaluated, suggesting that the students were developing an

Heidi Andrade is an Assistant Professor, Department of Educational and Counseling Psychology, University at Albany, 1400 Washington Avenue., ED 233, Albany, NY 12222; handrade@uamail.albany.edu. At the time of writing both YingDu andXiaolei Wang were PhD candidates in the Department of Educational and Counseling Psychology, University at Albany.

 Copyright © 2008 by the National Council on Measurement in Education

understanding of the qualities of effective writing as defined by the rubrics they received. Andrade concluded that simply handing out and explaining a rubric can increase students' knowledge of the criteria for writing but that translating that knowledge into actual writing is more demanding. She recommended sustained attention to the process of assessing writing, including involving students in the design of rubrics by critiquing sample pieces of writing, and by teaching students to self-assess their works in progress.

Andrade's recommendation regarding involving students in co-creating rubrics by critiquing examples is supported by research on the power of models in promoting skill acquisition. Zhu, Simon, and colleagues (Zhu, Lee, Simon, & Zhu, 1996; Zhu & Simon, 1987; Zhu, Zhu, Lee, & Simon, 2003) have demonstrated that studying worked-out examples of science or math problems can help students acquire new information and skills, use the skills to solve new problems, and express solutions efficiently and accurately. Wiggins (1998) argues that examples or models can be equally useful in teaching writing. Noting that the performance standards on rubrics are open to interpretation and that some students' views of "what it means to meet these criteria and the standard may be way off the mark" (p. 183), Wiggins recommends giving students models in order to promote more accurate analyses of the criteria in a rubric. Orsmond, Merry, and Callaghan (2004) agree that a key factor in self-assessment is students' understanding of specific criteria, and recommend the use of a subject-specific exemplar.

For these reasons, students in the treatment group in this study were given a model essay or story and asked to generate a list of criteria for their writing assignments by listing the qualities that made the model effective. Because we needed to use very similar or identical rubrics in different classes in order to make cross-class comparisons, students were not involved in co-creating entire rubrics. Rather, they were asked to generate a list of the criteria for their assignment, which invariably matched the rubrics that they were given during the next class.

The second quasi-experimental study we found also supports the belief that rubrics and self-assessment can promote learning. In a study of group learning in five sixth grade social studies classes, Cohen, Lotan, Scarloss, Schultz, and Abram (2002) found that students who were informed of the evaluation criteria for written essays had higher-quality discussions and better group products than students who worked without knowing the criteria. Using path analysis, these authors concluded that knowledge of evaluative criteria had an indirect, not a direct, effect on essay scores, with group products and self-assessment (group discussions of the quality of their product) playing a key mediating role.

Several studies of student responses to rubrics provide indirect support for the assertion that rubrics provide a learning advantage. In a qualitative study of undergraduates (Andrade & Du, 2005), students reported that they actively used rubrics to support their learning and academic performance. In focus groups, students discussed the ways in which they used rubrics to plan an approach to an assignment, check their work, and guide or reflect on feedback from others. They said that using rubrics helped them focus their efforts, produce work of higher quality, earn better grades, and feel less anxious about an assignment. Studies by Holmes and Smith (2003) and by Orsmond, Merry, and Callaghan (2004) have had similar results: Students reported increased clarity regarding an assignment when they have "criterion-referenced schemes" (Orsmond, Merry, & Callaghan, 2004, p. 275) and students believed that rubric-referenced assessment was more fair and "valuable to their learning" (Holmes & Smith, 2003, p. 320).

Research by Schafer, Swanson, Bené, and Newberry (2001) also lends indirect support to the view of students as users of assessments. Their study of the effects of teacher knowledge of rubrics on student achievement suggests that coaching high school algebra and biology teachers on the rubric used to evaluate student work on constructed response test items was associated with higher scores on tests. Schafer et al. speculate that the higher test scores are the result of teachers incorporating operational definitions of achievement into their instruction in ways that were understood and used by students.

Student Self-Assessment

Like rubric-referenced assessment, student self-assessment and its potential to increase learning and achievement have been receiving a lot of attention, much of it rhetorical (e.g., Baldwin, 2004; Leonhardt, 2005). The empirical evidence is mounting and compelling, however. White and Frederiksen (1995, 1998) were among the first to have shown success in using self-assessment in their study of a mechanics curriculum for junior high school students. More recently, a study of seventh and eighth grade students' writing by Andrade and Boulay (2003) found a relationship between self-assessment and quality of writing, especially for girls. Ross, Rolheiser, and Hogaboam-Gray (1999) have also studied self-assessment and writing. They report that weak writers in fourth, fifth, and sixth grade who were trained in self-assessment of narrative writing outperformed weak writers in the comparison group. Ross et al. note that changes in the conventions of language (sentence structure, grammar, and spelling) were negligible: The higher posttest scores of the weakest writers were the result of stronger performance on plot development, or the "integration of story elements around a central theme" and "the adoption of a narrative voice" (p. 124).

Ross, Hogaboam-Gray, and Rolheiser (2002) tell of similarly positive results of research involving self-assessment of word problems in fifth and sixth grade mathematics classes. Working with older mathematics students, Stallings and Tascione (1996) found that the process of self-assessment can "engage students in evaluating their progress, aid in developing their communication skills, and increase their mathematics vocabulary" (p. 548). Black, Harrison, Lee, Marshall, and Wiliam's (2003, 2004) study of formative assessment practices in math and science classes for 11- to 15-year olds revealed a strong relationship between formative assessment, including self-assessment, and achievement. These authors concluded that "the development of self-assessment by the student might have to be an important feature of any programme of formative assessment" (2003, p. 14). Finally, there is some evidence that self-assessment is associated with improved performance on traditional external examinations. A study of the influence of self-assessment training on high school students' scores on external exams (MacDonald & Boud, 2003) showed a consistently positive effect of treatment.

Research Questions and Hypotheses

The studies discussed above suggest a link between rubric-referenced self-assessment and student achievement. Given the recent surge in the popularity of rubrics in K-12 schools, however, there is a need for more research, particularly at the early elementary and high school levels. This study was designed to test popular claims about the relationship between rubric-referenced assessment, including and especially self-assessment, and elementary school students' writing by addressing two research questions:

1. Is there a main effect of a model, generating criteria, and rubric-referenced self-assessment on scores assigned to students' writing?
2. If so, is that effect influenced by gender, previous achievement in English/Language Arts, amount of time spent on writing an assignment in class, and/or prior exposure to rubrics?

We predicted the following hypotheses would be supported by our data:

1. Reading a model paper, generating the criteria for a rubric, and using a rubric to self-assess first drafts will be associated with higher scores for students' written work.
2. On average, girls will receive higher scores for their writing than boys.
3. Amount of writing time and previous achievement in English/Language Arts will be positively associated with writing scores.
4. Prior exposure to rubrics will not be related to writing scores.

Method and Data Sources

Participants

The study employed a convenience sample of 116 volunteer participants. It took place in seven public elementary school classrooms in the northeastern United States. Three classes were grade 3 ($N = 46$) and four were grade 4 ($N = 70$). Though each class had only one teacher and a teacher's aide for all academic subjects, five of the seven classes participated in the study during English/Language Arts (ELA) class time, and two of the classes participated in the study during social studies class time. The sample consisted of intact classes, matched by grade level and subject matter: There were four classes in the treatment group (two third grade classes studying ELA, and two fourth grades, one studying ELA and one studying social studies), and three in the comparison group (one third grade studying ELA, and two fourth grades, one studying ELA and one studying social studies).

Among the participants, 52 (44.8%) were boys, and 64 (55.2%) were girls. A majority of the participants ($N = 110$; 94.8%) were Caucasian. The number of participants who were identified as having special needs was very small ($N = 3$).

Instruments

Essays

Each class was asked to do a writing assignment. See Appendix A for each assignment. Six classes wrote persuasive essays. Four of the six classes wrote persuasive essays about year-round schooling, and two of the six—one treatment, one comparison—wrote about the impact of the European settlers on the Native Americans. One third grade treatment class wrote stories about their families. Writing about topics related to the curriculum was a condition of participation in the study imposed by the teachers of those classes.

Six researchers in three pairs scored the essays. The scorers were blind to treatment condition. The rubrics used in the treatment classrooms were adapted for use as scoring rubrics. Two scoring rubrics were developed: One for the third grade stories, and one for the third and fourth grade persuasive essays. The rubrics (Appendix B) were very similar but varied in terms of one criterion; the ideas and content of the writing to be assessed. In order to increase discrimination between levels and more precisely measure quality, the scoring rubrics included six levels of quality rather than the four levels in the rubrics used in the classrooms.

The scoring rubrics were tested by having the team of six researchers score a series of essays together. The rubrics were repeatedly revised until the scorers agreed that there was minimal ambiguity. A scoring procedure was designed to control individual scoring behaviors and to promote acceptable inter-rater reliability (Appendix C). The standards for inter-rater reliability were high: Rather than comparing total scores given by each rater, which can mask disagreements about the particular strengths and weaknesses of an essay, the scores given to each of the seven criteria for each essay were compared. The following scoring process was designed to manage the scoring done by each pair of scorers and to track inter-rater reliability:

1. *Anchor papers.* For each new grade level and/or essay topic, each pair of researchers scored two or three anchor essays together and reached consensus on the scores. Of the essays scored, 11% were co-scored as anchor papers.
2. *Co-scoring.* An additional 23% of the essays were co-scored by a pair of researchers. Each scorer scored the essays independently and shared her scores and her reasons for them with her co-scorer. They then came to agreement on final scores. Inter-rater reliability for co-scoring (that is, prior to the sharing of scores) was tracked for these essays. Percentage agreement ranged from 26% to 57% for complete agreement ($M = 43.3\%$), from 49% to 82% for scores that differed by half a point or less ($M = 67.2\%$), and from 81% to 97% for scores that differed by one point or less ($M = 90.5\%$). The first author scored nearly all of the anchor papers and co-scored essays in order to check for scoring drift.
3. *Final inter-rater reliability.* Another 14% of the essays were scored independently by two scorers in order to calculate the percent of final inter-rater reliability. Unlike the co-scoring step above, scores for these essays were not discussed. Averages of the scores assigned by the pair were entered into the database. The first author scored these essays as well, in order to guard against scoring drift. If she believed that the scores were not sufficiently grounded in the scoring rubric, the scoring pair was asked to co-score additional essays in order to promote agreement. This was rare. Final inter-rater reliability ranged from 21% to 71% for complete agreement ($M = 40.5\%$), from 64% to 100% for scores that differed by half a point or less ($M = 74.7\%$), and from 86% to 100% for scores that differed by one point or less ($M = 96.2\%$).
4. *Independent scoring.* The remaining 52% of the essays were divided between team members and scored independently.

Previous achievement in English/language arts

In order to include previous achievement in writing as a covariate in the analysis, recent English/Language Arts grades for reading, spelling, and writing were summed. Because grading scales differed by teacher (some assigned letter grades while others assigned numerical grades), all class grades were transformed into numbers (A+ = 97, A = 92, and so on). Previous achievement scores ranged from 209 to 297. The treatment and comparison groups had comparable means ($M = 267.49$ and $M = 266.04$, respectively).

Prior exposure to rubric use

Students' recent previous exposure to rubrics was measured via two questions on a questionnaire administered at the beginning of the study: "Has your teacher for *this class* ever given you a rubric for a writing assignment? (Yes or No.) If yes, about how many times has your teacher given you a rubric for a writing assignment? (1–2 times, 3–5 times, 6–10 times, 10 or more times.)"

Procedures

The writing process in each class resembled a Writers' Workshop: Students engaged in some form of prewriting, wrote first drafts, got feedback from the classroom teacher, and wrote final drafts. Table 1 summarizes the sequence of events followed by each class. In order to ensure fidelity of treatment, the first author co-led class periods one, two, and four with the classroom teachers.

The treatment condition differed from the comparison condition in three ways: The students in the treatment group (1) read a model story or essay, discussed its strengths and weaknesses, and generated a list of qualities of an effective story or essay; (2) received a written rubric; and (3) used the rubric to self-assess their first drafts. The students in the comparison group did not read a model story or essay but did generate a list of qualities of an effective story or essay. The comparison group did not receive a rubric. Students in the comparison group were asked to review their first drafts and note possibilities for improvement in the final draft. They did not self-assess their drafts according to a rubric.

Model stories or essays

The treatment group was given a model story or essay. The first author read the model aloud and asked students to critique it in terms of its strengths and weaknesses. Once they had soundly critiqued the model the students were asked to generalize by listing the criteria for their own written assignments. Their brainstormed list of criteria was tracked on the blackboard. Students were told that their list of criteria would be included in the rubric they would receive during the next class, and it was, as students always identified the major characteristics of effective writing. For the purposes of research, however, the rubrics given to the classes in the treatment group were very similar to each other and identical to those in Appendix B; students did not co-create idiosyncratic rubrics.

Rubrics

The rubrics given to the treatment group classes referred to seven commonly assessed criteria for writing (e.g., the 6 + 1 Trait® Writing Method; see Culham, 2003; Spandel & Stiggins, 1997): ideas and content, organization, voice and tone, word choice, sentence fluency, and conventions (see Appendix B). We treated the essay score data as interval-level data for analysis.

Self-assessment

During the self-assessment done in the treatment group, students were asked to underline key phrases in the rubric with colored pencils (e.g., "clearly states an opinion"), then underline or circle in their drafts the evidence of having met the standard articulated by the phrase (e.g., his or her opinion). If they found they had not met the standard, they were asked to write themselves a reminder to make improvements when they wrote their final drafts. This process was followed for each criterion on the rubric except the

Table 1. Sequence of Events by Group and Class Period

Group	Class Period 1	Class Period 2	Class Period 3	Class Period 4	Class Period(s) 5+
Treatment	1. Introduce Written Assignment	1. Hand out and Discuss Written Rubric	Students Write First Drafts	Students Use Rubric to Self-Assess First Drafts	1. Classroom Teacher Gives Each Student Feedback
	2. Read and Discuss Model Story or Essay	2. Prewriting, e.g., Outlining, Brainstorming			2. Students Write Final Drafts
	3. Generate List of Qualities of an Effective Story or Essay				
Comparison	1. Introduce Written Assignment	Prewriting, e.g., Outlining, Brainstorming	Students Write First Drafts	Students Self-Assess Drafts Without Rubric	1. Classroom Teacher Gives Each Student Feedback
	2. Generate List of Qualities of an Effective Story or Essay				2. Students Write Final Drafts

conventions criterion, which was not formally self-assessed.

Class time to write

Students were given class time to complete each step of the writing process. Time spent on writing—not instruction nor the treatment—was recorded in minutes by the first author.

Results

The average time spent on writing was 170.0 minutes ($SD = 49.6$). The amount of class time devoted to writing varied by class, from 105 to 265 minutes. The average number of minutes to write was 187.5 for the treatment group and 141.7 for the comparison group. Writing time was not significantly correlated with essay score ($r = .15, p = .1$), and a regression analysis indicated a statistically insignificant F-test ($F = 2.8$, $p = .1$).

Class averages of students' responses to the questions about prior rubric use by their teachers ranged from 0 (no, my teacher has not given me a rubric for a writing assignment) to 2.53 (yes, my teacher has given me a rubric for a writing assignment either 1–2 or 3–5 times) for the treatment group ($M = .99$), and from .31 to 1.26 for the comparison group ($M = .92$). Prior rubric use was not significantly correlated with writing score ($r = -.17$, $p = .07$), and a t-test indicated group equivalence in terms of previous rubric use ($t = 1.3$, $p = .24$). Regression analysis indicated an insignificant F-test for student-reported prior rubric use ($F = 3.24, p = .07$).

The average writing score for the entire sample was 26.8 ($SD = 5.2$), with a range of 16 to 40. In order to check for an effect of writing assignment topic on scores, we examined the mean scores represented in Table 2. The mean of the third grade stories ($M = 31.05$) was higher than the means of the other assignments but the stories were written by a treatment group, so we predicted higher scores. However, it is likely that third graders found it easier to write effective stories based on personal experience than to write persuasive essays, and that the familiarity of the story genre contributed to the higher mean. For this reason we analyzed the data both with and without the scores for the stories.

The means in Table 2 suggest no clear advantage to either the year round school or European-Native American essays. A t-test of the mean difference between the year round school essays and the two other assignments combined (stories and European-Native American essays) was not significant ($t = -1.8, p = .08$).

Table 3 contains the means and standard deviations for total essay score by treatment condition and gender. Initial analysis and data screening suggested that the variables most likely to influence students' writing scores include treatment, gender, and previous achievement in English/Language Arts. Ethnicity and special needs were not included as variables because the sample sizes for each are very small.

Total Essay Scores

Controlling for previous achievement, a GLM two-way ANOVA was used to analyze the main effect of treatment and gender. The assessment is statistically significant. On average, the treatment group's writing scores ($M = 28.5, SD = 4.9$) are higher than the comparison group's scores ($M = 24.3$, $SD = 4.7$), $F(1, 111) = 18.9$, $p = .000$, partial $\eta^2 = .15$. Girls tended to have somewhat higher essay scores ($M = 27.7$, $SD = 5.1$) than boys ($M = 25.8$, $SD = 5.2$), but the difference does not reach statistical significance, $F(1, 111) = 1.9$, $p = .17$. The interaction between treatment and gender is not significant, $F(1, 111) = .84$, $p = .36$. Figure 1 presents the estimated marginal means of total essays scores by treatment condition and gender.

ANOVA was also used to analyze the main effect of treatment and gender without the story scores ($N = 97$). The assessment is still statistically significant. On average, the treatment group's writing scores ($M = 27.5$, $SD = 4.6$) are higher than the comparison group's scores ($M = 24.3, SD = 4.7$), $F(1, 92) = 10.08$, $p = .002$, partial $\eta^2 = .10$. Again, no statistically significant gender difference is found, $F(1, 92) = 1.8$, $p = .18$. The interaction between treatment and gender is also not significant, $F(1, 92) = .84, p = .36$.

Scores on Individual Criteria

The previous paragraph reported on total essay scores, which were the sum of scores received for the seven criteria on the rubrics—ideas and content, organization, paragraph formatting, voice and tone, word choice, sentence fluency, and conventions. We also examined the relationships between the treatment and particular aspects of writing, as represented by each criterion. Table 4 contains the means and standard deviations for scores on each criterion on the rubric, by condition and gender. The main effects of the variables on the scores for individual criteria were examined using a GLM multivariate test. The results show that, controlling for previous achievement, treatment has a statistically significant relationship with criteria scores. For treatment, $F(7, 105) = 4.2$, $p = .001$, partial $\eta^2 = .22$. The effect of gender is not significant $F(7, 105) = .95$, $p = .47$. The interaction between treatment

Table 2. Means and Standard Deviations of Total Essay Scores by Grade Level, Treatment Condition, and Writing Assignment Topic ($N = 116$)

Grade	Condition	Topic	Mean	SD	N
3	Comparison	Year-Round School	25.0	3.66	13
	Treatment	Story	31.05	4.74	19
	Treatment	Year-Round School	26.39	4.85	14
4	Comparison	Year-Round School	26.15	4.57	15
	Comparison	Euro-Native American	22.32	4.91	18
	Treatment	Year-Round School	26.22	4.57	17
	Treatment	Euro-Native American	29.35	4.0	20

Table 3. Means and Standard Deviations of Total Essay Scores by Condition and Gender (*N* = 116)

Condition	Gender	Mean	SD	*N*
Comparison	Male	23.92	5.22	23
	Female	24.73	4.18	23
	Total	24.33	4.69	46
Treatment	Male	27.25	4.77	29
	Female	29.32	4.82	41
	Total	28.46	4.87	70
Total	Male	25.78	5.20	52
	Female	27.67	5.07	64
	Total	26.82	5.19	116

and gender is not significant F (7, 105) = 1.4, p = .23.

The between-subject effect tests shows that treatment has a statistically significant relationship with all criteria scores at the $p < .01$ level (Ideas, F = 20.7, p = .00, partial η^2 = .16; Organization F = 15.9, p = .00, partial η^2 = .13; Paragraphs F = 6.7, p = .01, partial η^2 = .06; Voice F = 11.2, p = .001, partial η^2 = .09; Word Choice F = 11.4, p = .001, partial η^2 = .09) except for Sentences (F = 2.9, p = .09) and Conventions (F = .48, p = .49). Gender is not significantly related to criteria scores.

Discussion

This study provides support for the hypothesis that having students use model papers to generate criteria for a writing assignment and using a rubric to self-assess first drafts is positively related to the quality of their writing. The treatment has a statistically significant, positive association with third and fourth grade students' essay scores, even controlling for the predictably powerful effect of previous achievement in English/Language Arts. The effect size for total essay scores (partial η^2 = .15) is small but meaningful in practice: Roughly translated into typical classroom grades (an admittedly subjective process that can be undertaken in a variety of ways) by equating a score of six on each criterion with 100%, a five on each criterion with 90%, a four with 80% and so on, the average grade for the treatment group would be a low B, compared to the average comparison group grade of a high C.

The influence of gender on writing scores was relatively predictable: We found that girls tended to receive somewhat higher scores for their writing than boys, though the differences were not statistically significant. The influence of amount of time spent on writing was, perhaps surprisingly, not related to total scores. We attribute this finding to the teachers' expert judgments of whether or not their students needed more time to work on their assignments.

Student-reported prior rubric use was not associated with total scores on the writing assignments. We attribute this finding to minimal rubric use by most of the students prior to the study. A poll of the teachers revealed that, of the seven teachers in the study, six had already exposed their students to rubrics but only four handed their rubrics out to students before they began an assignment, and only two had asked students to use the rubrics to assess their own or each others' work. Like Andrade (2001), we conclude that simply handing out a rubric is unlikely to have a measurable effect on student work.

In the analysis of the scores received on individual criteria (ideas and content, organization, paragraph formatting, voice and tone, word choice, sentence structure, and conventions), the treatment had a significant influence on all criteria except for sentences and conventions. Interestingly, conventions was the only criterion not explicitly attended to during the rubric-referenced self-assessment done by the treatment group. We interpret this finding as additional evidence of the effect of formal self-assessment.

We also note that the finding regarding the effect of treatment on criteria such as voice, or ideas and content, stands as a rejoinder to recent critiques of rubrics (Kohn, 2006; Wilson, 2006). Kohn and Wilson argue that rubrics promote weak writing by focusing attention on only the most quantifiable and least important qualities of assignments. The fact that rubric-referenced self-assessment was associated with higher scores on important qualities like ideas and content testifies to the potential of such processes to help students master significant, meaningful aspects of writing—at least when the rubrics emphasize those important qualities and when students are actively involved in using them (Andrade, 2006), as in this study.

There are several limitations to this investigation. One is the short treatment time: Students in this study were asked to write only one assignment. Since teachers who use rubrics often use them more than once, research on the long term effects of rubric use by students could be illuminating. Another is the use of only one, strong model paper. Critiquing multiple samples, including a weak paper, could

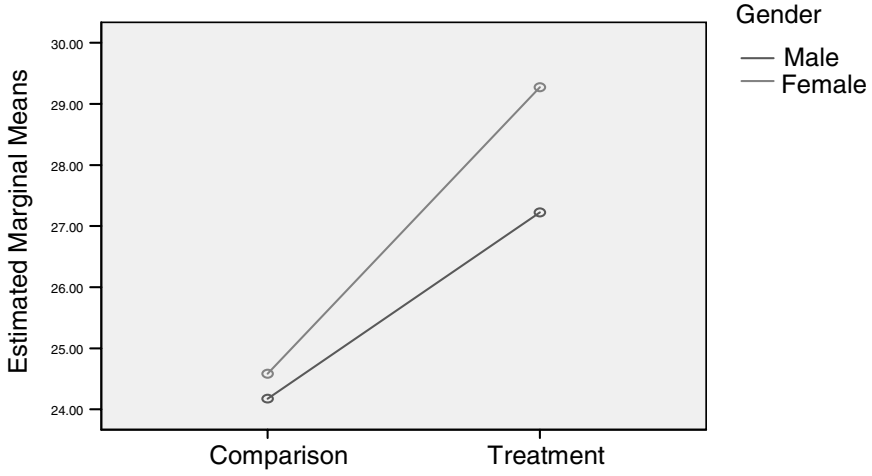

FIGURE 1. Estimated marginal means of total essays scores by condition and gender.

Table 4. Means and Standard Deviations of Scores for Each Criterion, by Condition and Gender ($N = 116$)

	Condition	Gender	Mean	*SD*	*N*
Ideas	Comparison	Male	3.02	1.32	23
		Female	2.76	1.42	23
		Total	2.89	1.36	46
	Treatment	Male	3.72	1.29	29
		Female	4.26	1.10	41
		Total	4.04	1.21	70
Organization	Comparison	Male	2.86	1.07	23
		Female	3.01	0.90	23
		Total	2.93	0.98	46
	Treatment	Male	3.61	0.94	29
		Female	3.76	1.00	41
		Total	3.70	0.97	70
Paragraphs	Comparison	Male	2.64	1.37	23
		Female	2.90	0.99	23
		Total	2.77	1.19	46
	Treatment	Male	3.46	1.42	29
		Female	3.44	1.45	41
		Total	3.45	1.42	70
Voice	Comparison	Male	3.70	1.55	23
		Female	4.03	1.20	23
		Total	3.86	1.38	46
	Treatment	Male	4.34	0.87	29
		Female	4.85	0.98	41
		Total	4.64	0.96	70
Words	Comparison	Male	3.59	0.85	23
		Female	3.97	0.81	23
		Total	3.78	0.84	46
	Treatment	Male	4.29	1.03	29
		Female	4.46	0.94	41
		Total	4.39	0.97	70
Sentences	Comparison	Male	4.17	0.70	23
		Female	4.10	0.67	23
		Total	4.14	0.68	46
	Treatment	Male	4.27	0.52	29
		Female	4.43	0.66	41
		Total	4.36	0.61	70
Conventions	Comparison	Male	3.95	0.72	23
		Female	3.96	0.89	23
		Total	3.95	0.80	46
	Treatment	Male	3.56	0.81	29
		Female	4.12	1.00	41
		Total	3.89	0.96	70

reduce the risk of students simply reproducing the model. Though we did not have a "cookie cutter" problem with the writing students did for this study, we know that good instructional practice involves the provision of multiple models.

Other limitations are related to common issues with classroom-based research: The lack of random assignment to treatment or comparison groups, which could have led to nonequivalent groups; multiple teachers with varying teaching styles, which makes it difficult to parse the effect of teacher and the effect of treatment; and the different writing assignments, which adds an additional source of variation between the groups. Though we attempted to manage some of these methodological issues by controlling for previous achievement in our statistical analyses and ensuring fidelity of treatment, we acknowledge the limitations inherent in the design of this study.

The implications for classroom practice that emerge from this research seem relatively straightforward: Elementary school students ought to be actively engaged in critiquing sample pieces of writing, in thinking together about the criteria that are in the rubrics by which their writing will be evaluated, and in self-assessment of their works in progress. In classrooms unconstrained by the demands of research, the process of co-creating criteria can even lead to student-generated rubrics. By involving students in the assessment process in these ways, teachers can blur the distinction between instruction and assessment and transform classroom assessment into a moment of learning (Zessoules & Gardner, 1991).

We limited our recommendations for classroom practice to the teaching of writing because that is what we studied. Studies like this one are needed in other domains, including and especially science and math, which tend to involve students in qualitatively different kinds of work. We also encourage research on rubric-referenced assessment in secondary schools and higher education, with diverse populations, and with students with learning disabilities.

Acknowledgments

The first author is grateful to the State of New York/United University Professions Joint Labor-Management Committee for supporting this research with a Dr. Nuala McGann Drescher Leave Award. The authors also thank Robin Akawi, Susan Rogers and Vicki Rosenberg for their assistance with the research.

References

Andrade, H. G. (2000). Using rubrics to promote thinking and learning. *Educational Leadership, 57*(5), 13–18.

Andrade, H. G. (2001, April 18). The effects of instructional rubrics on learning to write. *Current Issues in Education, 4*(4). Available at http://cie.ed.asu.edu/volume4/number4.

Andrade, H. (2006). The trouble with a narrow view of rubrics. *English Journal, 95*(6), 9.
Andrade, H., & Boulay, B. (2003). The role of rubric-referenced self-assessment in learning to write. *Journal of Educational Research, 97*(1), 21–34.
Andrade, H., & Du, Y. (2005). Student perspectives on rubric-referenced assessment. *Practical Assessment, Research and Evaluation, 10*(3), 1–11.
Arter, J., & McTighe, J. (2001). *Scoring rubrics in the classroom: Using performance criteria for assessing and improving student performance*. Thousand Oaks, CA: Corwin Press.
Baldwin, D. (2004). A guide to standardized writing assessment. *Educational Leadership, 62*(2), 72–75.
Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2003). *Assessment for learning: Putting it into practice.* Maidenhead, England: Open University Press.
Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2004). Working inside the black box: Assessment for learning in the classroom. *Phi Delta Kappan, 86*(1), 9–21.
Black, P., & Wiliam, D. (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan, 80*(2), 139–148.
Brookhart, S. (2003). Developing measurement theory for classroom assessment purposes and uses. *Educational Measurement: Issues and Practice, 22*(4), 5–12.
Cohen, E., Lotan, R., Scarloss, B., Schultz, S., & Abram, P. (2002). Can groups learn? *Teachers College Record, 104*(6), 1045–1068.
Culham, R. (2003). *6 + 1 traits of writing: The complete guide, grades 3 and up*. Portland, OR: Northwest Regional Educational Laboratory.
Holmes, L., & Smith, L. (2003). Student evaluations of faculty grading methods. *Journal of Education for Business, 78*(6), 318–323.
Kohn, A. (2006). The trouble with rubrics. *The English Journal, 95*(4), 12–15.
Leonhardt, A. (2005). Using rubrics as an assessment tool in your classroom. *General Music Today, 19*(1), 10–16.
MacDonald, B., & Boud, D. (2003). The impact of self-assessment on achievement: The effects of self-assessment training on performance in external examinations. *Assessment in Education, 10*(2), 209–220.
Moskal, B., & Leydens, J. (2000). Scoring rubric development: Validity and reliability. *Practical Assessment, Research and Evaluation, 7*(10). Available online: http://pareonline.net/getvn.asp?v=7&n=10.
Orsmond, P., Merry, S., & Callaghan, A. (2004). Implementation of a formative assessment model incorporating peer and self-assessment. *Innovations in Education and Teaching International, 4*(3), 273–290.
Popham, J. W. (1997). What's wrong—and what's right—with rubrics. *Educational Leadership, 55*(2), 72–75.
Quinlan, A. (2006). *Assessment made easy: Scoring rubrics for teachers from K-college.* Lanham, MD: Rowman and Littlefield Education.
Ross, J., Rolheiser, C., & Hogaboam-Gray, A. (1999). Effects of self-evaluation training on narrative writing. *Assessing Writing, 6*(1), 107–132.
Ross, J. A., Hogaboam-Gray, A., & Rolheiser, C. (2002). Student self-evaluation in grade 5–6 mathematics effects on problem-solving achievement. *Educational Assessment, 8*(1), 43–59.
Schafer, W., Swanson, G., Bené, N., & Newberry, G. (2001). Effects of teacher knowledge of rubrics on student achievement in four content areas. *Applied Measurement in Education, 14*(2), 151–170.
Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher, 29*(7), 4–14.
Spandel, V., & Stiggins, R. J. (1997). *Creating writers: Linking writing assessment and instruction* (2nd ed.). New York: Longman.
Stallings, V., & Tascione, C. (1996). Student self-assessment and self-evaluation. *Mathematics Teacher, 89*(7), 548–555.
Stiggins, R. J. (2001). *Student-involved classroom assessment* (3rd ed.). Upper Saddle River, NJ: Merrill/Prentice-Hall.
Stix, A. (1996, November). *Creating rubrics through negotiable contracting and assessment*. Paper presented at the National Middle School Conference, Baltimore, MD.
White, B. Y., & Frederiksen, J. R. (1995). *Developing metacognitive knowledge and processes: The key to making scientific inquiry and modeling accessible to all students*. Technical report no. cm-95-04. Berkeley: School of Education, University of California at Berkeley.
White, B. Y., & Frederiksen, J. R. (1998). Inquiry, modeling and meta-cognition: Making science accessible to all students. *Cognition and Instruction, 16*(1), 3–118.
Wiggins, G. (1998). *Educative assessment: Designing assessments to inform and improve student performance*. San Francisco, CA: Jossey-Bass Publishers.
Wilson, M. (2006). *Rethinking rubrics in writing assessment*. Portsmouth, NH: Heinemann.
Zessoules, R., & Gardner, H. (1991). Authentic assessment: Beyond the buzzword and into the classroom. In V. Perrone (Ed.), *Expanding student assessment*. Alexandria, VA: ASCD.
Zhu, X., Lee, Y.,Simon, H. A., & Zhu, D. (1996). Cue recognition and cue elaboration in learning from examples. *Proceedings of the National Academy of Science, 93*(3), 1346–1352.
Zhu, X., & Simon, H. A. (1987). Learning mathematics from examples and by doing. *Cognition and Instruction, 4*, 137–166.
Zhu, X., Zhu, D., Lee Y., & Simon, H. A. (2003). Cognitive theory to guide curriculum design for learning from examples and by doing. *Journal of Computers in Mathematics and Science Teaching, 22*, 285–322.

Appendix A: Writing Assignments

Third Grade Family Story Writing Assignment

Families seem to love telling stories about things that happened to them or their relatives. For example, one family tells about the time their grandmother finished sewing the hem of her daughter's wedding dress just as the wedding march music began to play and the doors opened for the bride. Another family often tells the story of their 4-year old son, who was told to "remember to look up" when reciting a poem at a school program. The boy proudly recited his poem while looking straight up at the ceiling!

Please write a family story. You may or may not have lived it but you have probably heard the story told so many times that you can retell it in detail. Be sure to write about what happened, to whom, and why the story is loved by your family.

Third and Fourth Grade Persuasive Essay Assignment: Year Round Schools

Most schools in America are open 10 months of the year, and closed for 2 months (July and August). This schedule was created when many Americans were farmers and children were needed to work on farms during the summer. Some people argue that we should move to year-round schools now that times have changed. Year-round schools would still be open only 180 days a year but the schedule would be spread out differently. For example, students could be in school for 9 weeks and off for 3 weeks all year long, including summer. The normal breaks for weekends and holidays would be built into this calendar.

Arguments *for* Year-Round School

- Students forget a lot during the summer. Shorter vacations can help them remember more.
- It is a waste of space and time to have school buildings closed all summer.
- Short breaks give students time to learn in other places besides schools.
- Students who need extra help can get it during the short breaks.
- Students get bored during the long break of summer.
- It is easier to schedule vacations because not everyone wants to travel at the same time.
- Other countries have year-round schools.

Arguments *against* Year-Round School

- Studies of year-round school have not always shown that it helps students learn more.
- Students will forget information whether they are out of school for three weeks or ten weeks.
- Teachers will have to review four times a year instead of just once.
- Summer programs such as camps will suffer.
- Students won't be able to get summer jobs.
- Many schools are old and do not have air conditioning.
- Band and sports could have problems with scheduling practices and competitions.
- Parents could have children at different schools on different schedules.

Please write 3 paragraphs in which you argue for or against year-round school. Be sure to clearly state your opinion, either "I want year-round school" or "I do not want year-round school." Give at least 3 reasons in support of your opinion.

Fourth Grade Persuasive Essay Assignment: Impact of Europeans on Native Americans

Europeans brought many changes to the Native Americans when they came to America. Some of the changes they brought were:

- The idea of owning land
- The idea of naming places
- Goods to trade such as cloth, knives, axes, guns, tools, pots and pans, and other metals
- New diseases
- New types of buildings.

Some people say these changes had a positive (good) impact on the lives of the Native Americans. Other people disagree, saying the changes had a negative (bad) impact on the Native American way of life. Please take a side. Did the changes brought by the Europeans have a positive or negative impact on the lives of the Native Americans?

Appendix B. Rubrics Given to Students in the Treatment Condition Family Story Rubric (3rd grade)

	3	2	1	0
Ideas and Content	The paper tells an interesting family story with lots of relevant details. It tells why the family loves the story. It stays on topic.	The paper tells a story but without much detail. Does not tell why family loves the story. Stays on topic.	There is a very general story but the writing strays off topic or doesn't give enough detail.	The story is unclear. It may be repetitious or disconnected thoughts with no main point.
Organization	The story has a beginning with an interesting lead, a middle, and an ending. It is in an order that makes sense.	The story has a beginning, middle and end. The order makes sense.	The story has an attempt at a beginning and or ending. Some ideas seem out of order.	There is no real beginning or ending. Ideas seem loosely strung together.
Voice	The writer sounds like a real person who likes the story. The story tells what the people in it thought and felt.	The writer seems sincere but not enthusiastic. Tells a little about what people thought and felt.	The story could have been written by anyone. Tells very little about what people thought and felt.	The writing is bland. It sounds like the writer doesn't like the story. No thoughts or feelings.
Word Choice	Descriptive words are used ("comical" instead of "funny" or "miserable" instead of "sad").	The words are mostly ordinary, with a few attempts at descriptive words.	The words are ordinary but generally correct.	The same words are used over and over. Some words are used incorrectly.
Sentence Fluency	The sentences are complete, clear, and begin in different ways.	The sentences are usually correct.	There are many incomplete sentences and run-ons.	The story is hard to read because of incomplete and run-on sentences.
Conventions	Spelling, punctuation, capitalization, grammar and paragraphs are correct. Only minor edits are needed.	Spelling, punctuation and caps are usually correct. Some problems with grammar and paragraphs.	There are enough errors to make the writing hard to read and understand.	There are so many errors that the writing is almost impossible to read.

Persuasive Essay Rubric (Grades 3 & 4)

	4	3	2	1
Ideas and content	The paper clearly states an opinion and gives 3 clear, detailed reasons in support of it.	An opinion is given. One reason may be unclear or lack detail.	An opinion is given. The reasons given tend to be weak or inaccurate. May get off topic.	The opinion and support for it is buried, confused and/or unclear.
Organization	The paper has a beginning with an interesting lead, a middle, and an ending. It is in an order that makes sense. Paragraphs are indented and have topic and closing sentences and main ideas.	The paper has a beginning, middle and end. The order makes sense. Paragraphs are indented; some have topic and closing sentences.	The paper has an attempt at a beginning and or ending. Some ideas may seem out of order. Some problems with paragraphs.	There is no real beginning or ending. The ideas seem loosely strung together. No paragraph formatting.
Voice & tone	The writing shows what the writer thinks and feels. It sounds like the writer cares about the topic.	The writing seems sincere but not enthusiastic. The writer's voice fades in and out.	The paper could have been written by anyone. It shows very little about what the writer thought and felt.	The writing is bland and sounds like the writer doesn't like the topic. No thoughts or feelings.
Word choice	Descriptive words are used ("helpful" instead of "good" or "destructive" instead of "bad").	The words are mostly ordinary, with a few attempts at descriptive words.	The words are ordinary but generally correct.	The same words are used over and over. Some words are used incorrectly.
Sentence fluency	The sentences are complete, clear, and begin in different ways.	The sentences are usually correct.	There are many incomplete sentences and run-ons.	The story is hard to read because of incomplete and run-on sentences.
Conventions	Spelling, punctuation, capitalization, and grammar are correct. Only minor edits are needed.	Spelling, punctuation and caps are usually correct. Some problems with grammar.	There are enough errors to make the writing hard to read and understand.	The writing is almost impossible to read because of errors.

Appendix C: Essay Scoring Procedures for Essay Raters

Scoring dozens of essays can be grueling, and the temptation to hurry is strong. The consequences of hurrying can be severe though, and include lots more time spent on getting inter-rater reliability, distorted research findings, and me wringing hands and necks, to name a few. That is why I am strict about my scoring procedures, and I will ask you to be strict as well. Please be sure you are well and rested, give yourselves frequent breaks, and use this process every time you score:

1. Get a glass of water and get comfortable. Breathe deeply. (Your brain works better when it is hydrated and oxygenated!)
2. Check to see if the essays you are about to score are on the list of essays NOT to score.
3. Read the assignment. Note the highlighted sections and make a mental note to keep an eye out for the content they specify.
4. Read two benchmark essays and their scored rubrics to remind you of the scores they received.
5. Read a new essay.
6. Score one criterion at a time. E.g., read the *entire* Ideas and Content criterion, from highest to lowest levels. Chose a score (you can overlap levels if necessary). Then do the Organization criterion the same way etc. Don't forget to read all 6 levels of each criterion as you score. Incredibly, you'll find you forget what is in each box, and your accuracy will suffer if you don't constantly remind yourself of the scoring possibilities by reading an entire criterion/row.
7. To score Word Choice, make squiggly lines under every descriptive ("big") word that is used correctly. Put a squiggly circle around incorrectly used big words. Count them up. Note repeated uses of "boring" words like bad, mad, sad, nice, happy.
8. To score variety in the beginnings of sentences for Sentence Fluency, circle the first word in each sentence. Count the number of times each word starts a sentence and

write a tally in the margin of the paper. Count run-ons and fragments.

9. To score Conventions, highlight incorrect spellings, capitalization and punctuation. Count them. Count grammar problems, too. Note which, if any, are due to the use of fancy techniques, such as dialogue.
10. When you finish an essay, read it again and check your scores for each criterion. Change them as necessary.
11. Look up. Take a drink of water. Smile. Breathe. Think about what a joy it is to be reading these essays. Start another one.