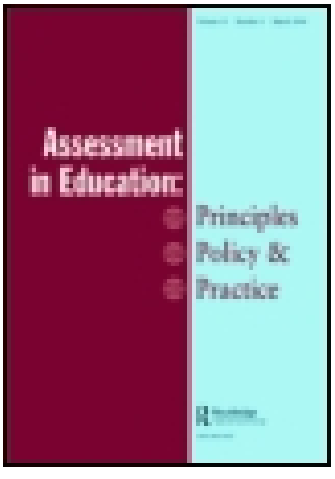

Assessment in Education: Principles, Policy & Practice

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/caie20

Rubric-referenced self-assessment and middle school students' writing

Heidi L. Andrade [a], Ying Du [b] & Kristina Mycek [a]

[a] Educational Psychology and Methodology, University at Albany, Albany, New York, USA

[b] American Board of Pediatrics, Chapel Hill, North Carolina, USA

Published online: 30 Apr 2010.

To cite this article: Heidi L. Andrade, Ying Du & Kristina Mycek (2010) Rubric-referenced self-assessment and middle school students' writing, Assessment in Education: Principles, Policy & Practice, 17:2, 199-214, DOI: 10.1080/09695941003696172

To link to this article: http://dx.doi.org/10.1080/09695941003696172

Assessment in Education: Principles, Policy & Practice
Vol. 17, No. 2, May 2010, 199–214

Routledge
Taylor & Francis Group

Rubric-referenced self-assessment and middle school students' writing

Heidi L. Andrade[a], Ying Du[b] and Kristina Mycek[a]

[a]*Educational Psychology and Methodology, University at Albany, Albany, New York, USA;*
[b]*American Board of Pediatrics, Chapel Hill, North Carolina, USA*

This study investigated the relationship between middle school students' scores for a written assignment ($N = 162$) and a process that involved students in generating criteria and self-assessing with a rubric. Gender, time spent writing, grade level, prior rubric use, and previous achievement in English were also examined. The treatment involved using a model essay to scaffold the process of generating a list of criteria for an effective essay, reviewing a written rubric, and using the rubric to self-assess first drafts. The comparison condition involved generating a list of criteria and reviewing first drafts. Findings include a main effect of treatment, gender, grade level, writing time, and previous achievement on total essay scores, as well as main effects on scores for every criterion on the scoring rubric. The results suggested that reading a model, generating criteria, and using a rubric to self-assess can help middle school students produce more effective writing.

Research has shown that feedback tends to promote learning and achievement (Bangert-Drowns et al. 1991; Butler and Winne 1995; Crooks 1988; Hattie and Timperley 2007), yet most students get little informative feedback on their work (Black and Wiliam 1998). The scarcity of feedback in most classrooms is due, in large part, to the fact that few teachers have the luxury of being able to respond regularly to each student's work. Fortunately, research also shows that students themselves can be useful sources of feedback via self-assessment (Andrade and Boulay 2003; Andrade, Du, and Wang 2008; Ross, Rolheiser, and Hogaboam-Gray 1999).

Self-assessment

Self-assessment is a process of formative assessment during which students reflect on the quality of their work, judge the degree to which it reflects explicitly stated goals or criteria, and revise accordingly. The emphasis here is on the word *formative*. Self-assessment is done on drafts of works in progress in order to inform revision and improvement: it is not a matter of having students determining their own grades. Given what we know about human nature, as well as findings from research regarding students' tendency to inflate self-evaluations when they will count toward formal grades (Boud and Falchikov 1989), we subscribe to a purely formative type of student self-assessment.

There is some empirical evidence that actively involving students in using a rubric to self-assess their writing is associated with noticeable improvements in achievement.

*Corresponding author. Email: handrade@uamail.albany.edu

ISSN 0969-594X print/ISSN 1465-329X online

DOI: 10.1080/09695941003696172
http://www.informaworld.com

For example, a study of seventh and eighth grade students' writing by Andrade and Boulay (2003) found a positive relationship between self-assessment and quality of writing, especially for girls. Ross, Rolheiser, and Hogaboam-Gray (1999) have reported that weak writers in fourth, fifth, and sixth grades who were trained in self-assessment of narrative writing outperformed weak writers in the comparison group. They note that changes in conventions of language (sentence structure, grammar, and spelling) were negligible: the higher post-test scores of the weakest writers were the result of stronger performance on substantive criteria such as plot development.

Andrade, Du, and Wang (2008) also looked at the effectiveness of rubric-referenced self-assessment on scores on elementary school (grades three and four) students' writing. Their findings indicated that having students use model papers to generate criteria for a writing assignment and use a rubric to self-assess first drafts is positively related to the quality of their subsequent writing. Like Ross and his colleagues (1999), Andrade, Du, and Wang found that the improvements in students' writing included more effective handling of sophisticated qualities of writing such as ideas and content, organisation, and voice.

Rubrics as self-assessment tools

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; Jonsson and Svingby 2007; Moskal 2003; Popham 1997). 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 and McTighe 2001; Quinlan 2006; Spandel 2006; Stiggins 2001). 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' (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), Shepard (2000), Brookhart (2003), 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 three relevant quasi-experimental studies; two of these studies also involved students in some form of self-assessment or self-grading. In a study of group learning in five sixth grade social studies classes, Cohen et al. (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. In a study of rubric-referenced self-grading by seventh grade students of constructed-response biology test items, Sadler and Good (2006) found significant and large gains in post-test scores, particularly for students whose pre-test scores were in the lower third of the range.

In a study of the relationship between seventh and eighth grade students' writing and simply giving the students a rubric when their assignment was introduced at the outset of the lesson (Andrade 2001), having a rubric was associated with higher scores on only one of the three essays written by the students for the study. 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 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 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 of high quality writing has indirect support from research on the power of models in promoting skill acquisition in science and math. Zhu, Simon, and colleagues (Zhu et al. 1996; Zhu and Simon 1987; Zhu et al. 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 also be 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' (183), Wiggins recommends giving students models of effective writing 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 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.

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 who had engaged in regular criterion-referenced self-assessment of their work for a course in educational psychology (Andrade and 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. Other studies of undergraduates have had similar results: students reported increased clarity regarding an assignment when they have 'criterion-referenced schemes' (Orsmond, Merry, and Callaghan 2004, 275) and students believed that rubric-referenced assessment was more fair and 'valuable to their learning' (Holmes and Smith 2003, 320).

Research by Schafer et al. (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. For this reason, prior rubric use by the teachers in this study was measured and analysed.

Research questions and hypotheses

The research described here was conducted to extend the research briefly reviewed above by engaging students in grades five through seven in a process of rubric-referenced self-assessment, including co-creating the criteria for an essay by discussing a model essay, and using a rubric to self-assess drafts of writing assignments. This study was designed to test popular claims about the relationship between rubric-referenced self-assessment and middle school students' writing by addressing two research questions:

(1) Is there a main effect of rubric-referenced self-assessment on scores assigned to students' writing?
(2) If so, is that effect mediated by gender, grade level, time spent on writing, and/or previous achievement in English?

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

(1) The treatment – 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) The amount of writing time will be positively associated with writing scores.
(4) Previous achievement in English/Language Arts will be positively associated with writing scores.

Method and data sources

Participants

The study employed a convenience sample of 162 students in 11 classes taught by six teachers. Five classes were in public schools, and six were in a private school for girls, all in the northeastern United States. Grade levels included grades five, six, and seven. Nine of the 11 classes were English/Language Arts; the two seventh grade groups, one treatment and one comparison, were History/Social Studies classes. The majority of the participants (86%) were Caucasian, 75% were girls, and 8% were students identified with special needs.

The sample consisted of intact classes. Assignments to the treatment or comparison group were made in terms of two variables: grade level and the degree to which the teachers had already used rubrics with their classes. Information about teachers' prior rubric use was collected via conversations with the teachers and brief classroom observations. Five classes ($N = 89$) were assigned to the treatment group, and six to the comparison group ($N = 73$).

Instruments

Essays

Each class wrote a persuasive essay. See Appendix 1 for each assignment. Seven of the classes wrote about year-round schools. A relatively new development in the United States and elsewhere, year-round schools are typically in session the same number of days per year as schools on a traditional calendar, but the two-month long summer break is distributed over the year; see Appendix 1. Four classes (two treatment, two comparison) wrote about either child labor laws (fifth grade) or the dropping of the atom bomb on Japan (seventh grade). Writing about topics related to the curriculum was a condition of participation in the study imposed by the teachers of those classes.

Rubrics

The rubric given to the treatment group referred to six commonly assessed criteria for writing (e.g., the 6+1 Trait® Writing Method; see Spandel and Stiggins 1997): ideas and content, organisation, voice and tone, word choice, sentence fluency and conventions. A seventh criterion related to paragraph formatting was also included at the request of some of the teachers, who noted that their students needed to pay special attention to this aspect of organisation (see Appendix 2). The four gradations of quality were written at a vocabulary level that is appropriate for middle school students, and in language generic enough to be applied to different topics. In order to assess criterion validity (or external validity; Messick 1996), or the relationship between assessment scores and other measures, the essay scores obtained in this study were compared to the students' English/Language Arts standardised test scores, when available. Criterion validity was evidenced by significant correlations between the standardised test scores and total essay scores ($r = .4$, $p < .001$). We treated the essay score data as interval level data for analysis.

A team of researchers, blind to experimental condition, participated in scoring the essays. The rubric used in the treatment was adapted for use as a scoring rubric (see Appendix 2). In order to increase discrimination between levels and more precisely measure quality, the scoring rubric included six levels of quality rather than the four levels in the rubric used in the treatment.

The scoring rubrics were tested by having the team of 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. 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. Eight percent ($N = 13$) of the essays were co-scored as anchor papers, 39% ($N = 63$)

were co-scored by two scorers, 17% ($N = 27$) were independently scored by two scorers, and the remaining 36% of the essays ($N = 59$) were scored independently. In addition, the principal investigator scored all of the anchor essays, and most of the co-scored and twice-scored essays in order to guard against scoring drift. Final inter-rater reliability was 38% complete agreement, 73% for scores that differed by half a point or less, and 94% for scores that differed by one point or less on the six-point scale.

Previous achievement in English/Language Arts

In order to include previous achievement in writing as a covariate in the analyses, the state-administered, standardised English/Language Arts (ELA) test scores and the most recent ELA class grades given by the teachers in the public schools were collected. The standardised ELA tests for public middle schools consist of a section containing multiple-choice and short- or extended-response questions based on reading selections, and a section containing multiple-choice and short- or extended-response questions based on a listening selection. The tests given in grades five and seven also include an editing task. The tests given in grades six and eight contain short-response and extended-response questions based on paired reading selections.

For the private school, only the teacher-assigned English class grades were available for most of the students. In order to produce an adequate sample size for analysis, the ELA scores from the public school and the class grades from the private school were transformed into z-scores with a mean of 500 and a standard deviation of 100 (based on the reporting of College Entrance Examination Board scores).

Prior exposure to rubric use

In order to verify the balance between the comparison and treatment groups in terms of students' recent exposure to rubrics, the students were asked to respond to 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 pre-writing, wrote rough drafts, got feedback from the classroom teacher, and wrote final drafts. The treatment condition differed from the comparison condition in three ways: the students in the treatment group (1) read a model essay, discussed its strengths and weaknesses, and generated a list of qualities of an effective 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 essay or receive a rubric but they did generate a list of qualities of an effective essay. Students in the comparison group were also asked to review their first drafts and note possibilities for improvement in the final draft (see Table 1).

In order to ensure fidelity of treatment, the first author co-led class periods one, two, and four with the classroom teachers. The participating teachers did not receive training before the study began.

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 essay 2. Read and discuss model essay 3. Generate list of qualities of an effective essay	1. Hand out and discuss written rubric 2. Pre-writing, e.g., outlining, brainstorming	Students write first drafts	Students use rubric to self-assess first drafts	1. Classroom teacher gives students feedback 2. Students write final drafts
Comparison	1. Introduce essay 2. Generate list of qualities of an effective essay	Pre-writing, e.g., outlining, brainstorming	Students write first drafts	Students self-assess drafts without rubric	1. Classroom teacher gives students feedback 2. Students write final drafts

Model stories or essays used to co-generate criteria

The treatment group was given a model 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 generalise 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, since 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 2; students did not co-create idiosyncratic rubrics.

Self-assessment

During the self-assessment done in the treatment group, students were asked to underline key phrases in the rubric with coloured pencils (e.g., 'clearly states an opinion'), then underline or circle in their drafts the evidence of having met the standards articulated by the phrases (e.g., his or her opinion). If they found they had not met the standards, 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 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, at the discretion of the classroom teacher. Time spent on writing – not instruction or the treatment – was recorded in minutes by the first author.

Results

The average time spent on writing was 133.3 minutes ($SD = 37.3$). The amount of class time devoted to writing varied by class, from 75 to 175 minutes. The average number of minutes to write was 112.4 for the treatment group and 158.7 for the comparison group. Writing time was not significantly correlated with essay score ($r = .09$, $p = .24$), and a regression analysis indicated a statistically non-significant F-test ($F = 1.38$, $p = .24$). Writing time was included in subsequent analyses anyway, however, given the well-established relationship between time on task and achievement (Brophy 1988).

Class averages of students' responses to the questions about prior rubric use by their teachers ranged from .38 (indicating the students tended to recall being given a rubric for a writing assignment between zero times and once) to 3.28 (indicating the students tended to recall being given a rubric for a writing assignment three to five times) for the treatment group ($M = 2.1$), and from 0 to 2.6 for the comparison group ($M = 1.23$). Although a t-test indicated group non-equivalence in terms of previous rubric use ($t = 6.1$, $p < .01$), rubric use was not significantly correlated with writing score ($r = .14$, $p = .1$), and regression analysis indicated a non-significant F-test ($F = 3.34$, $p = .07$). Prior rubric use was not included in further analyses because it is a class level variable, and also because it was unlikely to have a measurable relationship with essay scores given the relative infrequency of rubric use by teachers in both groups.

A poll of the teachers revealed that, of the six teachers in the study, five had shared their rubrics with students but only two asked students to actively use the rubrics to assess their own or each other's work.

Total Essay Scores

Total Essay Score is the combination of scores received for the seven criteria on the scoring rubric: ideas and content, organisation, paragraph formatting, voice and tone, word choice, sentence fluency, and conventions. The maximum possible total essay score is 42. The average score for the entire sample was 29.0 ($SD = 4.8$), with a range of 16 to 39. In order to check for an effect of writing assignment topic on scores (year round schools, child labor laws, or the dropping of the atom bomb on Japan) we examined the means and standard deviations of Total Essay Scores by grade level, treatment condition, and writing assignment topic. The mean of the seventh grade (both a treatment and a control group) essays about the atom bomb ($M = 26.99$) was lower than the means of the other assignments. However, a *t*-test of the mean difference between the year-round school essays and the two other assignments combined (atom bomb and child labor law; $N = 19$ and 25, respectively) was not significant ($t = .54$, $p = .59$). Essay topic was not included in further analysis.

Initial analysis and data screening suggested that the variables related to students' writing scores include treatment, gender, grade level, 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.

Controlling for previous achievement and writing time, a GLM two-way ANOVA was used to analyse the main effect of treatment, grade level, and gender. The assessment was statistically significant. On average, the treatment group's writing scores ($M = 30.4$, $SD = 4.7$) are higher than the comparison group's scores ($M = 27.4$, $SD = 4.3$), $F(1, 161) = 30.6$, $p < .001$, partial $\eta^2 = .17$. Girls tended to have somewhat higher essay scores ($M = 29.3$, $SD = 4.7$) than boys ($M = 28.3$, $SD = 4.9$), but the effect was not statistically significant $F(1, 161) = 0.12$, $p = .73$. Grade level was statistically significant $F(2, 160) = 7.6$, $p = .001$. There were no significant interaction effects.

Scores on individual criteria

The previous section reported on Total Essay Scores. We also examined the relationships between the treatment and particular aspects of writing, as represented by each criterion.

The main effects of the variables on the scores for individual criteria were examined using a GLM multivariate test: Ideas+Paragraphs+Words+Voice+Organisation+Sentences+Conventions = Intercept+Treatment+Gender+Grade+Previous Achievement Standard Score+Writing Time. The results show that, controlling for previous achievement and writing time, treatment has a statistically significant relationship with criteria scores. For treatment, $F(7, 148) = 7.4$, $p < .001$, partial $\eta^2 = .26$. The effect of gender is not significant $F(7, 148) = 0.43$, $p = .89$, but the effect of grade level was significant $F(14, 298) = 7.8$, $p < .001$, $\eta^2 = .27$.

The treatment group's average scores were higher than the comparison group's average scores for all seven criteria. A between-subject effect test shows that treatment has a statistically significant relationship with each criterion: Ideas ($F = 16.2$, $p < .001$, partial $\eta^2 = .10$); Organisation ($F = 5.2$, $p = .02$, partial $\eta^2 = .03$); Paragraph Formatting ($F = 28.7$, $p < .0001$, partial $\eta^2 = .16$); Voice ($F = 5.6$, $p = .02$, partial η^2

= .04); Word Choice (F = 26.1, p < .001, partial η^2 = .15); Sentences (F = 9.1, p = .003, partial η^2 = .06); and Conventions (F = 8.4, p = .004, partial η^2 = .05).

Discussion

This study provides support for the hypothesis that having middle school students use model papers to generate criteria for a writing assignment and self-assess first drafts according to a rubric is positively related to the quality of their writing. The treatment has a statistically significant, positive association with fifth, sixth and seventh grade students' essay scores, even controlling for the predictably powerful effects of previous achievement in English/Language Arts and time spent writing. The effect size for total essay scores (partial η^2 = .17, ω^2 =.11) 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 comparison group would be 77%, or a high C, compared to the average treatment group grade of 83%, or a B.

The influence of gender on writing scores was relatively predictable: we found that girls tended to receive higher scores for their writing than boys, but the differences were not statistically significant. The lack of the expected magnitude of differences between scores received by boys and girls has at least two possible explanations. One explanation is the fact that our analysis controlled for previous achievement in English/Language Arts. Another explanation is related to the sample: the seventh grade students, who received the lowest average scores, were all girls. There were no seventh grade boys in the sample.

In the analysis of the scores received for the individual criteria (ideas and content, organisation, paragraph formatting, voice and tone, word choice, sentence structure, and conventions), the treatment had a statistically significant association with every criterion. Effect sizes were largest for paragraph formatting and word choice (partial η^2 = .16 and .15, respectively). The relationships between treatment and the criteria found in this study reflect previous research that also found an association between these variables in elementary school students' writing (Andrade, Du, and Wang 2008), except for conventions. We explained this finding regarding conventions in the elementary school study in terms of the fact that conventions were not formally self-assessed. However, this criterion was not formally self-assessed by the middle school writers either; yet an association exists. It is possible that the older students informally self-assessed and revised the mechanics of their writing but we do not have the data needed to support this claim.

There are several limitations to this investigation. One is the short treatment time: students in this study were asked to write and self-assess only one assignment. Research is needed on the longitudinal effects of treatment. Another limitation is the use of only one, strong model paper. Critiquing multiple samples, including a weak paper, could 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. A third limitation is the imbalance in terms of gender: because of difficulties we experienced while working in a private school for boys, we have far fewer boys in the study than we had hoped for. Although the number of boys is adequate for statistical analysis, a more gender-balanced sample is desirable.

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 non-equivalent 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 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: middle school students ought to be actively engaged in critiquing sample pieces of writing, in thinking about the criteria contained in the rubrics by which their writing will be evaluated, and especially in a process of careful self-assessment of their works in progress. In classrooms unconstrained by the demands of research, the process of co-creating criteria can 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 and 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. Finally, we encourage research that honors the importance of instructional sensitivity by more closely aligning instruction and assessment than was possible in this study (e.g., Niemi et al. 2007).

Acknowledgement

The authors would like to thank Hongyu Cheng and Tony Leonardi for their help with the scoring of essays, and the reviewers of this manuscript for their thoughtful suggestions for revision.

Notes on contributors

Heidi Andrade is an associate professor at the University of Albany, State University of New York. Her scholarship is focused on classroom assessment, particularly student self-assessment, and the self-regulation of learning.

Ying Du is a psychometrician at the American Board of Pediatrics. She received her PhD from the University of Albany, State University of New York, in 2007. Her research interests include professional licensure and certification examinations, research on the pediatric workforce, and psychometric methods.

Kristina Mycek, MS, CAS, is a doctoral student in the division of Educational Psychology and Methodology at the University at Albany. Her work focuses on evaluation, measurement, statistics, and assessment.

References

Andrade, H.G. 2000. Using rubrics to promote thinking and learning. *Educational Leadership* 57, no. 5: 13–8.

Andrade, H.G. 2001. The effects of instructional rubrics on learning to write. *Current Issues in Education* 44. http://cie.ed.asu.edu/volume4/number4.

Andrade, H., and B. Boulay. 2003. The role of rubric-referenced self-assessment in learning to write. *Journal of Educational Research* 97, no. 1: 21–34.
Andrade, H., and Y. Du. 2005. Student perspectives on rubric-referenced assessment. *Practical Assessment, Research and Evaluation* 10, no. 3: 1–11.
Andrade, H., Y. Du, and X. Wang. 2008. Putting rubrics to the test: The effect of a model, criteria generation, and rubric-referenced self-assessment on elementary school students' writing. *Educational Measurement: Issues and Practices* 27, no. 2: 3–13.
Arter, J., and J. McTighe. 2001. *Scoring rubrics in the classroom: Using performance criteria for assessing and improving student performance.* Thousand Oaks, CA: Corwin Press.
Bangert-Drowns, R.L., C. Kulik, J. Kulik, and M. Morgan. 1991. The instructional effect of feedback in test-like events. *Review of Education Research* 61, no. 2: 213–38.
Black, P., and D. Wiliam. 1998. Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan* 80, no. 2: 139–48.
Boud, D., and N. Falchikov. 1989. Quantitative studies of student self-assessment in higher education: A critical analysis of findings. *Higher Education* 18, no. 5: 529–49.
Brookhart, S. 2003. Developing measurement theory for classroom assessment purposes and uses. *Educational Measurement: Issues and Practice* 22, no. 4: 5–12.
Brophy, J. 1988. Research linking teacher behavior to student achievement: Potential implications for instruction of Chapter 1 students. *Educational Psychologist* 23, no. 3: 235–86.
Butler, D., and P. Winne. 1995. Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research* 65, no. 3: 245–81.
Cohen, E., R. Lotan, B. Scarloss, S. Schultz, and P. Abram. 2002. Can groups learn? *Teachers College Record* 104, no. 6: 1045–68.
Crooks, T. 1988. The impact of classroom evaluation practices on students. *Review of Educational Research* 58, no. 4: 438–81.
Hattie, J., and H. Timperley. 2007. The power of feedback. *Review of Educational Research* 77, no. 1: 81–112.
Holmes, L., and L. Smith. 2003. Student evaluations of faculty grading methods. *Journal of Education for Business* 78, no. 6: 318–23.
Jonsson, A., and G. Svingby. 2007. The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review* 2, no. 2: 130–44.
Messick, S. 1996. Validity of performance assessments. In *Technical issues in large-scale performance assessment,* ed. G. Philips, 1–18. Washington, DC: National Center for Education Statistics.
Moskal, B.M. 2003. Recommendations for developing classroom performance assessments and scoring rubrics. *Practical Assessment, Research and Evaluation* 8, no. 14. http://PAREonline.net/getvn.asp?v=8&n=14 (accessed 24 November 2003).
Niemi, D., J. Wang, D.H. Steinberg, E.L. Baker, and H. Wang. 2007. Instructional sensitivity of a complex language arts performance assessment. *Educational Assessment* 12, nos. 3/4: 215–37.
Orsmond, P., S. Merry, and A. Callaghan. 2004. Implementation of a formative assessment model incorporating peer and self-assessment. *Innovations in Education and Teaching International* 4, no. 3: 273–90.
Popham, J.W. 1997. What's wrong – and what's right – with rubrics. *Educational Leadership* 55, no. 2: 72–5.
Quinlan, A. 2006. *Assessment made easy: Scoring rubrics for teachers from K-college.* Lanham, MD: Rowman and Littlefield Education.
Ross, J., C. Rolheiser, and A. Hogaboam-Gray. 1999. Effects of self-evaluation training on narrative writing. *Assessing Writing* 6, no. 1: 107–32.
Sadler, P., and E. Good. 2006. The impact of self- and peer-grading on student learning. *Educational Assessment* 11, no. 1: 1–31.
Schafer, W., G. Swanson, N. Bené, and G. Newberry. 2001. Effects of teacher knowledge of rubrics on student achievement in four content areas. *Applied Measurement in Education* 14, no. 2: 151–70.
Shepard, L.A. 2000. The role of assessment in a learning culture. *Educational Researcher* 29, no. 7: 4–14.
Spandel, V. 2006. In defense of rubrics. *English Journal* 96, no. 1: 19–22.

Spandel, V., and R.J. Stiggins. 1997. *Creating writers: Linking writing assessment and instruction.* 2nd ed. New York, NY: Longman.

Stiggins, R.J. 2001. *Student-involved classroom assessment.* 3rd ed. Upper Saddle River, NJ: Merrill/Prentice-Hall.

Wiggins, G. 1998. *Educative assessment: Designing assessments to inform and improve student performance.* San Francisco, CA: Jossey-Bass Publishers.

Zessoules, R., and H. Gardner. 1991. Authentic assessment: Beyond the buzzword and into the classroom. In *Expanding student assessment,* ed. V. Perrone, 47–71. Alexandria, VA: ASCD.

Zhu, X., Y. Lee, H.A. Simon, and D. Zhu. 1996. Cue recognition and cue elaboration in learning from examples. *Proceedings of the National Academy of Sciences* 93, no. 3: 1346–51.

Zhu, X., and H.A. Simon. 1987. Learning mathematics from examples and by doing. *Cognition and Instruction* 4, no. 3: 137–66.

Zhu, X., D. Zhu, Y. Lee, and H.A. Simon. 2003. Cognitive theory to guide curriculum design for learning from examples and by doing. *Journal of Computers in Mathematics and Science Teaching* 22, no. 4: 285–322.

Appendix 1: Writing assignments

Fifth and Sixth Grade Persuasive Essay Assignment: Year Round Schools

Most schools in America are open 10 months of the year, and closed for two 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 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 5 paragraphs in which you argue for or against year round school. Be sure to clearly state your opinion, either 'I am for year round school' or 'I am against year round school.' Give at least 3 reasons in support of your opinion.

Fifth Grade Persuasive Essay Assignment: Child Labor Laws

Child labor was common in the U.S. in the past and is still common in many places around the world. Although many countries have strict child labor laws, others do not. Some people argue that child labor should be outlawed so children are not forced to work long hours in terrible conditions, often far away from their families, for little or no money. Other people argue that developing countries need child labor in order to keep businesses and factories open and economies strong. Write a persuasive essay explaining why child labor should or should not be outlawed worldwide.

Seventh Grade Persuasive Essay Assignment: Atom Bomb

Write a 5 paragraph persuasive essay in which you argue for or against the dropping of the atomic bomb on Japan during World War II.

Appendix 2: Rubrics given to students in the treatment condition

Persuasive Essay Rubric (Grades 5 and 6, year round and child labor essays)

	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 essay 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.

Persuasive Essay Rubric (Grade 7, atom bomb essay)

	3	2	1	0
Ideas and Content	The topic is focused and main thesis is clear. Relevant, accurate facts & details provide evidence for the thesis. The author explains how the facts support the thesis, and addresses opposing views.	The topic is evident but broad and lacking in detail. The writing stays on topic but doesn't address minor parts of the assignment.	There is a very general topic but the writing strays off topic or doesn't address major parts of the assignment.	The topic and main ideas are unclear. The writing may be repetitious or disconnected thoughts with no main point.
Organization	Essay has an interesting motivator, developed middle, and scintillating conclusion that restates the thesis and blueprint in new words. There are at least 3 middle paragraphs with a topic sentence and concluding sentence. The order of ideas makes sense.	The paper has a beginning, middle and end. Sequencing is logical.	The paper has an attempt at an intro and/or conclusion. Some ideas seem out of order.	There is no real introduction or conclusion. Ideas seem strung together in a loose fashion.
Voice	The writing matches the purpose and audience. The author seems to care about the topic. Tone and style are engaging.	The writing seems sincere but the author's voice fades in and out.	The writer seems to be aware of an audience but does not attempt to engage it.	The writing is inappropriate for the purpose or audience, and bland or mechanical.
Word Choice	Uses specific, powerful words, striking phrases, and lively verbs.	Words used are adequate, with a few attempts at colourful language.	Words used are ordinary but generally correct.	Limited, repetitive vocabulary. Words are sometimes used incorrectly.
Sentence Fluency	Sentences are well-constructed and have different beginnings and lengths. Easy to read aloud.	Sentences are usually correct. Some variety in beginnings and length.	Many poorly constructed sentences. Little variety in beginnings/length.	The paper is hard to read because of incomplete, run-on and awkward sentences.
Conventions	Double spaced. Few errors in spelling, punctuation, capitalization, grammar. Numbers < 11 or starting sentences spelled out. Uses third person. No slang.	Conventions are usually correct. Some problems with grammar, syntax and/or paragraphing.	Errors are frequent enough to be distracting.	Frequent errors make the paper almost impossible to read.