Ross, J. A., Rolheiser, C., & Hogaboam-Gray, A. (1999). Effect of self-evaluation on narrative writing. *Assessing Writing*, 6(1), 107-132.

Effects of Self-Evaluation Training on Narrative Writing

John A. Ross* - Ontario Institute for Studies in Education
Carol Rolheiser - Ontario Institute for Studies in Education
Anne Hogaboam-Gray - Ontario Institute for Studies in Education

*Corresponding author
Dr. John A. Ross,
Professor & Head,
OISE/UT Trent Valley Centre,
Box 719, 150 O'Carroll Ave.,
Peterborough, ON
K9J 7A1 CANADA

Abstract

Despite the burgeoning use of alternate assessment, few studies have examined effects on students. In this study, 148 students in 15 grade 4-6 classrooms were taught over an 8-week period how to evaluate their work (control $N=148$). Treatment group students became more accurate in their self-evaluations than controls. Contrary to the beliefs of many students, parents and teachers, students' propensity to inflate grades decreased when teachers shared assessment responsibility. Treatment students also outperformed controls on narrative writing but the overall effect was small ($ES=.18$). Weaker writers improved their writing much more if they were in the treatment than the control group ($ES=.58$). Improvements consisted of increasing integration of story elements around a central theme and the adoption of a narrative voice. In contrast conventions of language were relatively unchanged. The results of the treatment were attributed to the focusing effect of joint criteria development and use, and to the heightened meaningfulness of self-evaluation over other assessment data.

Effects of Self-Evaluation Training on Narrative Writing¹

Teacher strategies for assessing students' work have shifted from an exclusive reliance on testing toward a more balanced approach in which classroom tests and examinations are supplemented with alternate forms such as portfolio assessment, performance evaluation, and self-evaluation. Assessment is more closely integrated with instruction, instruments and procedures are demystified, assessment is a continuous process rather than a terminal event, and teachers share authority with students.

Although self-evaluation has been implemented extensively in elementary schools, few systematic attempts to teach students how to evaluate their work have been reported and little is known about the effects of self-evaluation training on students' achievement. In this study we implemented an in-service program that provided a small sample of teachers with instruments and procedures for teaching grade 4-6 students how to evaluate their performance in narrative writing and measured the effects of the in-service on the accuracy of students' self-appraisals and the quality of their narrative writing.

Theoretical Framework

Motivation for Alternate Assessment Practices

Proponents of alternate assessment practices, defined as assessment that "occurs in motivating contexts with meaningful tasks that are a part of daily instruction" (van Kraayenoord & Paris, 1993, p. 524), claim that a shift from traditional to alternate assessment will increase student achievement (e.g., Stiggins, 1994; Wiggins, 1993). Learning is enhanced because alternate assessment encourages teachers to focus on the objectives to be measured and the assessments

provide teachers with more accurate information than traditional tests, enabling teachers to respond more precisely to students' learning needs.

There is consistent, though not extensive, evidence that alternate assessment influences teachers' instructional practices in positive ways (e.g., Dorfman, 1997; Khattri, Kane, & Reeve, 1995; Koretz, Stecher, Klein, & McCaffrey, 1994; Lipson & Mosenthal, 1997). Few studies have asked whether these changes increase student achievement. Studies of alternate assessment that attempted to answer questions about student outcomes produced mixed results. For example, Shepard, Flexer, Hierbert, Marion, Mayfield, and Weston (1996) found that performance assessment had a small positive effect on student achievement in mathematics but not reading. A state-wide portfolio assessment program was abandoned, partly because student scores on traditional and alternate evaluations declined (Chrispeels, 1997). There is some evidence that students' study habits change when performance assessment (Lee & Suen, 1995) or portfolio assessment (Slater, Ryan, & Samson, 1997) is introduced.

Rationale for Linking Self-Evaluation to Achievement

Our expectation that a self-evaluation assessment system would enhance student achievement was based on four arguments. Students will learn more because (i) self-evaluation will focus student attention on the objectives measured, (ii) students will pay more attention to the assessment, (iii) student motivation will be enhanced, and (iv) the assessment provides teachers with information they would otherwise lack.

Self-evaluation focuses student attention. It has long been demonstrated that being clear about goals makes a positive contribution to performance (Locke, Shaw, Saari, & Latham, 1981). Several studies (reviewed in Hillocks, 1986) found that the quality of student writing improved

when teachers explained the criteria on which student work would be judged. If students apply these criteria to assess their work, the effects should be even stronger (through rehearsal and focusing). Students should develop a clearer understanding of what they are supposed to do and how well they are doing it. In addition, students report that they have a better grasp of academic expectations when they are involved in setting the criteria on which their work will be judged (Ross, Rolheiser, & Hogaboam-Gray, 1998a).

These benefits are likely to accrue only if students are provided with outcome-based criteria at an appropriate level of generality. Not all self-evaluation procedures recommended for classroom use do so. For example, many of the instruments in Rhodes (1993) direct student attention to processes not outcomes: Students ask themselves whether they used each of the steps in the Writing Process, rather than how well they wrote. In contrast, Kulm (1994) recommends that teachers begin self-evaluation by involving students in the construction of scoring rubrics which students then use to appraise their work. But even the joint construction of rubrics might be insufficient to focus student attention if the rubrics are too task-specific (making it difficult for students to detect the underlying learning objectives), too general (simply an array of superlatives that fail to indicate what is essential in a quality response), or too complicated for students to use easily.

Students pay more attention to self-evaluation than to other assessments. As students move through the school system their skepticism about the validity of test scores increases (Paris, Lawton, Turner, & Roth, 1991), a trend that has also been observed in portfolio assessment projects (Paris, Turner, Muchmore, Perry, 1995). Students view self-evaluation more positively than other kinds of assessment. Ross et al. (1998a) found that students liked self-evaluation

because it increased clarity about expectations, was fairer, and gave students feedback they could use to improve the quality of their work. Students reported that with teacher feedback they focused on what they did well on or on the grade, whereas with self-evaluation they focused on what they needed to improve. But students need help in using evaluation data to set goals. The provision of valid performance data without goal setting support can decrease achievement for lower ability students (Fuchs, Fuchs, Karns, Hamlett, Katzaroff, & Dutka, 1997).

Student motivation will be enhanced. Self-evaluation has an indirect effect on achievement through self-efficacy (i.e., beliefs about one's ability to perform actions that lead to desired ends). The most powerful contributor to higher self-efficacy is mastery experience (Bandura, 1997), that is, students who have been successful in the past anticipate they will be successful in the future. But even unsatisfactory performance might not lead to depressed confidence if students believe that they could be successful by adopting a different strategy (Schunk, 1995). What is crucial is how students evaluate their performance. Positive self-evaluations encourage students to set higher goals and commit more personal resources to learning tasks (Bandura, 1997; Schunk, 1995). Negative self-evaluations lead students to embrace goal orientations that conflict with learning, select personal goals that are unrealistic, adopt learning strategies which are ineffective, exert low effort and make excuses for performance (Stipek, Recchia & McClintic, 1992). Higher self-efficacy translates into higher achievement (Pajares, 1996). Wagner (1991), in a linear structural model, found positive path coefficients from self-evaluation to self-efficacy and from self-efficacy to performance.

Self-evaluation provides teachers with information otherwise unobtainable. Conventional test procedures and many alternate performance tasks provide no information about students' inner states during task performance, their subsequent interpretations about the quality of their

work, and the goals they set in response to feedback. Self-evaluation is unique in asking students to reflect on their performance. Self-evaluation instruments that elicit information about students' effort, persistence, goal orientations, attributions for success and failure, and beliefs about their competence give teachers a fuller understanding of why students performed as they did. When incorporated into teachers' deliberative planning, data generated from self-evaluations enables teachers to present content and anticipate impediments to learning, especially motivational obstacles. But many self-evaluation procedures, particularly those in which students check categories of responses without elaboration, fail to provide such rich detail.

Student Skill in Evaluating Their Work

These arguments suggest that self-evaluation is potentially a powerful stimulant of achievement. But students harbor misconceptions about the process. For example, many do not appreciate the role that evidence plays in self-evaluation (Ross et al., 1998a) or how discrepancies between student and teacher appraisals are resolved (Ross, Rolheiser, & Hogaboam-Gray, 1999). Self-evaluation is unlikely to have a positive impact on achievement if these misconceptions are not addressed by teaching students how to assess their work.

Students' self-evaluations tend to be inflated. The self-evaluations of even young children correlate reasonably well with their teachers' appraisals when students are asked to make a global assessment, comparing their ability to that of their classmates (Crocker & Cheeseman, 1988). Accuracy is much lower for specific tasks, even for adults, if information about a specific ability is lacking or difficult to process (Bandura, 1997). Elementary students tend to over-estimate their success on school tasks, in part because they expect that teachers will give them tasks that they can complete (Schunk, 1996), but also because young students lack the cognitive skills required

to integrate information about their abilities and they are more vulnerable to wishful thinking.

Overestimates of specific performance are likely to lead to complacency and reduced effort. For example, the child who does not recognize the need for help will not seek it (Markman, 1979). Simply requiring self-evaluation is unlikely to have an effect on achievement. Students have to be taught how to evaluate their work accurately.

Results of Previous Attempts to Teach Self-evaluation

Few studies have examined the effects of teaching students how to self-evaluate in classroom settings over a sustained period (i.e., four weeks or more). Hillocks (1986) reviewed seven studies (all but one were unpublished dissertations) in which students were given scales for judging writing samples. Students used the scales to assess the writing of their peers, to give editing suggestions, and to re-write deficient passages. In some instances they evaluated their own writing, although in Hillocks' review self-evaluation was a minor theme in these studies. Positive effects on student performance were reported.

The only study of self-evaluation in the language area (Arter, Spandel, Culham, & Pollard, 1994) gave grade 5 students direct instruction on the meaning of six traits of essay writing. The teacher, without student participation, determined the traits. Students scored a sample of essays and applied trait analysis to their own writing over a five-month period. The treatment group outperformed controls on one of the six traits (ideas). But the analysis procedures failed to protect the findings from Type I error. The authors should have used multivariate analysis (to deal with the possibility of multicollinearity among the six dependent variables) or applied a Bonferroni adjustment (Serlin, 1993). If the latter had been applied (by dividing the alpha of $p < .05$ by six, the number of comparisons in the same set of dependent variables) none of the results would have been statistically significant.

No studies of the effects on student accuracy of teaching self-evaluation in elementary and secondary classrooms have been reported. Research on university students indicates that accuracy improves when professors and students agree on assessment criteria (Falchikov & Boud, 1989) and when students are asked to justify their assessments (Boud, Churches, & Smith, 1986). There is also evidence from short duration lab studies that the self-evaluation accuracy of elementary students can be improved by influencing goal conditions (Butler, 1990) and drawing attention to previous performance (Stipek, Roberts, & Sanborn, 1984).

Research Questions and Predictions

Our approach to teaching students how to evaluate their work began in a study of the student assessment practices of cooperative learning teachers nominated as exemplary by their peers and supervisors. We organized their strategies as a four-stage process: (i) involve students in defining evaluation criteria, (ii) teach students how to apply the criteria, (iii) give students feedback on their self-evaluations, and (iv) help students use evaluation data to develop action plans. Strategies for each stage were elaborated by a team of teachers and reported as a series of action research case studies and classroom usable tools (Rolheiser, 1996). Use of these strategies had a positive effect on student attitudes to evaluation in some but not all of the pilot test classrooms (Ross et al., 1998a; 1999). Our goal in the current study was to determine whether teaching students how to evaluate their work would improve achievement in narrative writing of students in grades 4-6. Our research questions and predictions were:

1. Will self-evaluation training increase the accuracy of students' self-assessments? We anticipated that students in the treatment group would evaluate their work more accurately because all four stages in our model reduce uncertainty about the criteria for judging academic work.

2. Will self-evaluation training contribute to language achievement? We anticipated that focusing student and teacher attention on performance criteria (Stages 1 and 2) would enhance achievement.

Method

The study was a quasi-experimental, pre-post equivalent groups design. Grade 4-6 teachers from one school district volunteered to teach self-evaluation in their language period (about 60 minutes per day). Similar teachers from an adjacent district were the control. Pretests (consisting of a narrative writing assignment and a battery of other measures) were administered at the beginning of the project. Teachers attended three in-service sessions and implemented strategies for teaching self-evaluation over an 8-week period. At the end of the study a posttest (consisting of a narrative writing assignment) was administered.

Sample

Students in the classrooms of 15 volunteer grade 4-6 teachers, in a large school district, constituted the treatment group. The control consisted of students in the classrooms of 15 volunteer teachers in an adjacent district. The teachers were matched on gender, grade, and academic training (as shown in Table 1). Within each class we randomly selected data from 10 students for analysis. In one treatment class only eight students obtained parental consent so we randomly deleted two students from one of the control classes. The total sample was 296 students.

Instruments

Students completed a battery of instruments at the beginning of the project in the following sequence: On Day 1 they completed a survey (described below) consisting of a self-efficacy measure (how sure they were they could write a good short story) and a locus of control measure. On Day 2 they wrote a short story. On Day 3 they evaluated their short story, shared their attitudes toward self-

evaluation and completed a goals orientation survey about their feelings when writing their short story. At the end of the project they repeated the writing and self-evaluation tasks.

Student Achievement. Students completed (a) a pre- and post-test narrative writing task. Teachers were asked to present the writing task in their usual manner. Teachers could have a class discussion of possible topics but they had to emphasize to students that it was an individual writing task. Teachers described the criteria on which the stories would be marked (plot or story development, characters, setting, providing interest for the reader and grammar/spelling). Students wrote a rough copy and then a final copy. We asked teachers not to provide editing help. There was little difference between rough and final copies, except in one treatment class in which the teacher edited some stories. In this class the rough copies were marked; in all other classes the final copies were coded.

We developed a six-level coding scheme (displayed in the appendix) by elaborating descriptions in the provincial writing rubric. Two anchor papers were identified for each level. Over a two-week period two teachers used the rubric to mark 592 stories. An English consultant who had been a trainer in provincial writing assessment programs trained the markers (by reviewing the six levels in terms of the anchor papers, marking and discussing additional papers from the pilot test) in the first two days. During the training period, the 6-level scheme was expanded by distinguishing high and low performance (e.g., level 2 and 2+) within levels 2-5. The two markers then independently graded each paper, assigning a holistic 1-6 score. After each set of ten papers the markers resolved discrepancies in their assessments through discussion. Before discussion the between-rater agreement was Cohen's $\kappa=.66$ for perfect agreement and $\kappa=.86$ for agreement within one half level on the six-point scale. The papers were marked in random order, intermingling pre- with post-tests and treatment with control

students. The markers and the trainer were blind to the experimental conditions of the students and to study goals.

(b) Accuracy of self-evaluation was calculated from the achievement data and from student responses to survey items administered immediately after the achievement task. Students used a 1-10 scale (anchored by 1=not well and 10= very well) to rate the quality of their story. They used the same 1-10 scale to answer five additional probes to rate how well they wrote each part: plot, characters, setting, interest for the reader, and grammar and spelling. These six items were averaged to create a 1-10 mean score for each student. The self-evaluation scores and the achievement scores were bifurcated at their medians and combined to create three groups: accurate (low self-evaluation with low achievement or high self-evaluation with high achievement), underestimate (low self-evaluation with high achievement) and overestimate (high self-evaluation with low achievement).

Student Instruments for Estimating Sample Equivalence. We estimated the pretest equivalence of the two groups with three measures. Since student goals influence learning (evidence reviewed in Urdan & Maehr, 1995), we administered a goals orientation survey. 15 items from Meece, Blumenfeld, and Hoyle (1988) distinguished three orientations toward learning tasks: mastery (e.g., “The work made me want to find out more about the topic.”), ego (e.g., “I wanted others to think I was smart.”), and affiliative (e.g., “I wanted to help others with their work.”). Other studies (reviewed in Pajares, 1996) indicate that student self-efficacy consistently predicts achievement. Our self-efficacy measure consisted of the same 6 items included in our self-evaluation instrument except that each asked “how sure are you that you could...” rather than “how sure are you that you [did]”. Since response to training in self-evaluation could be influenced by pre-existing student attitudes to self-evaluation we

administered a survey consisting of 10 Likert items adapted from Paris, Turner, and Lawton (1990) and Wiggins (1993).

Teacher Instruments for Estimating Sample Equivalence We also examined pretest equivalence by comparing the 15 teachers in each sample on two constructs linked to assessment practice or language achievement. Ten Likert items measured teachers' use of assessment methods that are fair, transparent, participatory, and collaborative (e.g., "My students help me interpret assessment results."). Teachers also completed 16 items from Gibson and Dembo (1984) measuring personal teaching efficacy (e.g., "When I really try, I can get through to even the most difficult students.") and general teaching efficacy (e.g., "The amount that a student can learn is primarily related to family background."). Both types of teacher efficacy correlate with teachers' willingness to try new ideas and student achievement (evidence reviewed in Ross, 1995b, 1998). Teachers also provided demographic information (e.g., gender, experience, certification).

Experimental Conditions

The contexts for all teachers were similar. The grade 4-6 language program in operation in both districts encouraged a whole language approach, emphasizing the writing process accompanied by skill development components to enhance conventions of language. The provincial curriculum specified learning expectations for each grade and specified standards for evaluation, for example, that it should be ongoing, collaborative, tied to learning outcomes, and directed toward the improvement of instruction. Self-evaluation was identified as a necessary and valuable skill in both provincial and school district policies.

Teachers in the treatment condition attended 3 three-hour, after-school in-service sessions distributed over the eight weeks of the field test. The in-service included discussion and

experiential activities focused on the following topics: 1) self-evaluation and why it is important; 2) a theoretical model of how self-evaluation contributes to learning; 3) a four-stage model for teaching self-evaluation to students using narrative writing (i.e., involving students in defining criteria, teaching them how to apply the criteria, giving students feedback on their self-evaluations, and helping them develop action plans); 4) the advantages and disadvantages of the four-stage model; 5) forms that could be used to teach self-evaluation; and, 6) participant successes and collaborative problem-solving. During the course of the treatment the teachers also attended four brief team meetings in their schools to review progress and solve problems that arose during their enactment of the treatment.

The in-service sessions and the handbook (Rolheiser, 1996) provided examples of lessons and practice activities that most teachers could adapt. Although some of these examples focused specifically on language development, most focused on assessing social skills. Teachers had complete control over how they adapted these materials to the language curriculum; however, the four-stage model provided a structure or process to guide them.

The in-class activities consisted of 4-30 minute lessons in which the teacher demonstrated a particular self-evaluation technique or engaged students in discussion of their self-evaluations. All treatment teachers reported at the in-service that they had involved students in defining the criteria that would be used to evaluate their writing. The most common technique was to have students brainstorm potential criteria. The teacher recorded suggestions on a flip chart, adding her own suggestions and grouping student ideas into appropriate categories. Students voted on which criteria were most important. Each of the selected criteria was given to a student group that identified levels of performance (e.g., high, medium, and low). The teacher, usually outside of

class, assembled student levels into a rubric for assessing performance. This rubric was progressively refined through the duration of the study. In these activities teachers shared control: they retained expectations for writing in provincial guidelines but used student language to express them and involved students in constructing performance descriptions that were at an appropriate level of generality.

Once a rubric or set of criteria had been established, teachers reported that they explained what each criterion meant. Many of the treatment teachers illustrated the levels with anchor papers (drawn from provincial or district documents). Teachers modeled the application of the criteria and students practiced applying the criteria to their own stories. An essential component was to give students feedback on the accuracy of their assessments. Most teachers reported conferencing with each student at least once during the study. Teachers reported that they asked students to set goals and actions for future writing based on their self-evaluations and feedback from the teacher. The in-class component also included 12 short practice sessions in which students completed a 3-5 minute self-evaluation using a form provided by the teacher.

The treatment differed from previous attempts to improve student writing by providing scales or assessment criteria. In this study, students contributed to the selection of the criteria and construction of the scales, the scales used student language, teachers provided scaffolding to ensure proper use, gave feedback on student understanding of the scales and helped students use the scales to develop goals.

During the 8 weeks of the project control group teachers continued teaching language as they usually did, including self-evaluation if that was part of their practice, but not emphasizing it. The assessment criteria that control condition teachers used were similar to those used in the

treatment because both sets of teachers were responsive to the provincial curriculum and provincial tests of language in grade 3 and 6 (annual school results were widely reported). But control teachers did not involve students in identifying criteria, did not provide overt guidance on how to self-evaluate, give feedback on the accuracy of self-appraisals, or help students set goals. Student work was evaluated by the teacher through grades and teacher comments.

To compensate for the attention given to the treatment group, control group teachers received a half-day of additional prep time to work on their writing curriculum. Treatment teachers received no release time.

Analysis

Descriptive statistics (means, standard deviations, reliabilities) for all student and teacher variables were compiled. Prior to inferential statistics all variables were normalized using log transformations. Pretest equivalence of the treatment and control groups was determined through a series of t-tests on all student and teacher variables. For the first research question, the proportion of students with accurate self-appraisals at the beginning and end of the project in the treatment and control groups were compared in contingency tables, using chi-square to determine statistical significance. For the second research question, an analysis of covariance was conducted in which the dependent variable was post-test achievement, the covariates were pre-test achievement and self-evaluation accuracy, and the independent variable was experimental condition.

Results

Equivalence of the Groups

Teacher Data Table 1 describes the teacher variables. The reliabilities of the three scales were acceptable. At the beginning of the project there were no significant differences between treatment and

control group teachers in terms of their use of alternate assessment practices, personal teaching efficacy, and general teaching efficacy. Table 1 also shows the teachers were similar in age, experience, gender and qualifications.

Table 1 About Here

Student Data. Table 2 summarizes the reliabilities (Cronbach's Alpha) of the student variables. The reliabilities for self-efficacy, self-evaluation, self-evaluation attitudes, and mastery goal orientations were adequate. The reliabilities for mastery and ego goal orientations were borderline.

Table 2 About Here

Table 3 summarizes the means and standard deviations of the student variables for each treatment condition. There was one pretest difference between the groups. Treatment students significantly outperformed control students on the pretest writing task [$t(290.164)=4.79, p=.022$]², a concern because the pretest writing task strongly predicted posttest writing scores [$r=.605, p<.001$] and posttest accuracy [$r=.384, p<.001$]. There was no other pretest student difference between the groups.

Table 3 About Here

Research Questions

The first research question asked whether training in self-evaluation increased the accuracy of student appraisals. There were no significant differences between treatment and control group students in self-evaluation accuracy on the pretest [$\chi^2(1, 284)=2.992, p=.084$]. On the post-test, treatment group students were significantly more accurate in their self-assessments [$\chi^2(1, 277)=7.037, p=.008$]. Table 4 shows the posttest accuracy rate within each experimental condition for three groups: students

who underestimated, overestimated, and accurately appraised their pretest performance. In each pretest group the accuracy rates were higher for treatment than control students but none of the differences reached statistical significance ($p < .05$).

Table 4 About Here

Table 5 shows the results of the analysis of covariance of student achievement (research question 2). The dependent variable was posttest writing score, the independent variable was experimental condition, and the covariates were pretest writing score and self-evaluation accuracy. Table 5 shows that only one of the covariates, pretest writing score was a significant predictor of achievement. Students who scored high on the pretest also scored high on the posttest as expected by the correlation ($r = .605$). The table shows that the correlation of pretest accuracy with post achievement ($r = -.245$) was spurious. Pretest achievement, which predicted post achievement, was one of the terms in the calculation of pretest accuracy. When pretest achievement was controlled, self-evaluation accuracy did not predict achievement.

Table 5 About Here

Both groups improved over the 8 weeks of the study. Table 5 shows that students who were taught how to evaluate their work wrote better narratives than students in the control group (after controlling for pretest differences between the groups). But the effect of the treatment was very small (treatment versus control $ES = .18$)³, accounting for only 2% of the variance compared to 28% for the pretest achievement covariate.

Table 5 also shows there was a treatment X pretest interaction. To understand it, we divided the sample into low (pretest scores 1-3) and high (pretest scores 4-6) achievers. Inspection of the cell means (after posttest achievement scores had been transformed) shows, in Table 6, that the treatment

had an impact only on low achievers. Students who produced poor writing samples on the pretest and were then taught how to evaluate their work substantially outperformed similar students in the control group ($ES=.58$). In contrast, students who wrote well on the pretest performed equally well on the posttest, regardless of whether they were given self-evaluation training.

Table 6 About Here

We re-read the writings of the lowest achievers in the treatment group, students who received 1, 2 or 2+ on the pretest ($N=23$). All but two of these students scored higher on the posttest with equal thirds gaining .5, 1.0, and 1.5 levels. Improvements were not equally distributed across all performance criteria. Changes in conventions of language (sentence structure, grammar, and spelling) were negligible and story lengths were virtually unchanged. The higher posttest scores of the weakest writers were the result of stronger performance on two dimensions.

First, virtually all students (86%) improved their treatment of narrative elements, particularly plot development. The posttest plots did not necessarily contain more events but they were organized more tightly around a central idea. For example, Courtney's pretest story consisted of loosely connected episodes about a visit to an island once occupied by pirates. Her posttest story focused on an experience with an intruder: the theme was introduced in the opening sentence (a noise in an adjacent room) and was elaborated in a series of connected events. Posttest plots were more likely to overtly involve the audience. For example, Andrew provided a flat description of an encounter between a man and a fortune teller in his pretest story. In his posttest story Andrew described how a boy awoke as a monster, evoking various reactions from his sister and friend; at the end of his story he asked readers why the boy's mother had not noticed

the change. Although the number of characters in the stories did not change in any consistent way, characters in posttest stories were more fully developed. For example, in her pretest story Sarah wrote about two space travelers that were undifferentiated: they had individual names but they experienced events together. Sarah's posttest story also contained two characters (a young girl and a Genie) but the protagonists were distinguished; their interaction, and dialogue, constituted the plot. For other students, improvements centered on the setting. For example, Brittany in her pretest story provided sparse information about the spacecraft on which events occurred but her posttest story contained details about the animals and vegetation of the rainforest.

The second area of improvement, involving 38% of the weakest writers, concerned voice. The protagonist in pretest stories produced by these students was the writer relating or fantasizing a personal experience in the first person. In posttest stories the child's narrative voice was beginning to emerge. For some students the main character was a child with the same name as themselves but the story was related from the perspective of the observer. Other students began writing in the third person although some reverted to the first half way through. Overall the writer became less visible in posttest stories, the events focused on characters expected to be of interest to the audience, and use of the third person increased.

Discussion

The first finding was that teaching self-evaluation skills increased the accuracy of student self-appraisals. The greatest impact was for students who were overestimating their performance, a sizeable proportion of the sample. This is an important finding for two reasons. First, students are unlikely to change how they go about writing narratives nor are they likely to seek help from teachers and peers, if they believe their work meets classroom standards. Second, students are

concerned about accuracy. Some students believe they lack the expertise to assess their work accurately; others anticipate that given an opportunity to have input to their grades, their peers will cheat (Ross et al., 1998a). Teachers have reservations about self-evaluation for the same reasons and teachers perceive parents to have similar fears (Ross, Rolheiser, & Hogaboam-Gray, 1998b). But in this study, when teachers shared assessment control with students the tendency to inflate grades decreased. Student involvement in rubric construction and receiving feedback on their application of the criteria gave students a clearer understanding of classroom standards. In addition, students were required to talk about the reasons for their self-evaluations, especially when there were discrepancies between a self-evaluation and the teacher's judgment. This focus on rubric-grounded evidence reduced the influence of other bases that students could use to award themselves grades, such as amount of effort and self-aggrandizement, that students rely on when standards are unarticulated.

The size of the impact of training on accuracy was small. One possible explanation might be that students participated in the development of classroom writing rubrics but did not contribute to the rubric on which their pre and posttest stories were graded.⁴ If they had seen the rubric, their judgments about how well they did might have been more accurate. Since the rubric was not available to them we do not know the basis for their self-evaluations. It could have been the rubrics for writing they had co-produced in class, an intuitive comparison between the piece they wrote for the research and pieces they had written previously or they may have been comparing their performance to the writing typically produced by their peers. An alternate explanation for the weak effects of the treatment on accuracy is that we focused on a learning objective that receives extensive instructional attention in every grade. Students might have been

so knowledgeable about what counts in writing that the focus on criteria and evidence contributed little to their understanding of what they were supposed to do. It might be that self-evaluation training would have a greater impact if focused on less familiar learning objectives.

The second finding was that self-evaluation training had a positive effect on achievement but only among weaker writers. The overall effect of the treatment was small ($ES=.18$), below the average effect size (.28) for the 75 writing treatments reviewed in Hillocks' (1986) meta-analysis. Part of the explanation for small overall effects might be the duration of the treatment (8 weeks). Hillocks found that treatments of less than 17 weeks had a lower mean effect (.21) than lengthier interventions. Greater duration may be required to overcome students' misconceptions about self-evaluation. In addition students need time to accept the idea that they have a role in assessment. In students' prior school experience evaluation was the teacher's exclusive prerogative. Another explanation is that the materials teachers used were not exclusively focused on writing, although only writing performance was measured in the study. The effects of self-evaluation training on writing skills may have been diluted when teachers used self-evaluation for social skills and in subjects other than language.

Self-evaluation had a much larger impact on the performance of students who wrote poorly at the beginning of the study ($ES=.58$). The reason might be that self-evaluation training gave poorer writers explicit feedback on what they needed to improve on that was more meaningful to them than the feedback they usually received from the teacher. In our previous studies of student cognitions about self-evaluation (Ross et al., 1998a; 1999) students reported paying more attention to self-evaluation because they understood the criteria, they felt ownership of the data, and they felt empowered because the teacher trusted them to rate themselves fairly.

Within-student differences in the stories indicated that the weakest writers improved in two areas if they were in the treatment group, integration of story elements around a central theme and the adoption of a narrative voice. In contrast conventions of language were relatively unchanged. Training in self-evaluation, especially participation in the development and application of assessment criteria, improved low achiever understanding of teacher expectations, shifting their attention from an exclusive focus on grammar and spelling to elements of the genre. But why did the treatment have minimal effect on higher achievers? It might be that the rubrics were oriented toward lower performers. It might also be that better writers knew what was expected of them and did not need the criteria to be spelled out in rubrics (at least for narrative writing).

The overall effect of self-evaluation on student achievement in this study was greater than has been reported for other alternate assessment measures. For example, Shepard et al. (1996) found that a project to introduce performance assessments (grade 3 teachers had weekly workshops for a year) had no overall effect on reading achievement and only a small effect on mathematics ($ES=.13$). The data from this study, particularly regarding the performance of poorer writers, presents a more encouraging picture.

Future research might build on these findings by, first, collecting observational data on how teachers use self-evaluation. In this study we relied on teacher self-reports of their enactment. In subsequent research researchers might (a) include treatment variation as an independent variable using a sample of observations based on the implementation profile in Rolheiser (1996) and (b) conduct qualitative case studies (our first is underway) that provide rich observational data on teachers' intentions and actions when teaching self-evaluation. Second, in this study we collected no data on student cognitions about self-evaluation, a theme we addressed

in a previous study (Ross et al., 1998a), that warrants further investigation. We need to understand better how self-evaluation differs from evaluation by the teacher in terms of how students use the instruments they are given (e.g., whether they rate their performance in terms of their past work, what other students do or in terms of an absolute scale), how they process the information (e.g., do they discuss the evaluation with others), and how they use self-evaluation data to set goals. Third, this study measured the effects of a compound treatment: (a) students were given explicit evaluation criteria and practiced applying them (as in some previous research), and (b) students participated in determining the criteria and in making decisions about the quality of their performance (a novel feature not previously investigated in studies of writing). Subsequent research might disentangle the relative impact of these two elements. Fourth, in our study we found only the weaker writers benefited from self-evaluation. We need to redesign our strategies for teaching self-evaluation for more able writers and measure the effects.

This study produced knowledge of two types. For researchers the study contributed evidence of the consequential validity of alternate assessment, a topic that has been seriously neglected despite recognition among test developers that the consequences of test use is a key factor in determining the worth of assessment instruments (Linn, 1997; Messick, 1995; Moss, 1992; Shepard, 1997). For teachers the study suggests that self-evaluation might be a useful mechanism for increasing student achievement and the accuracy of self-appraisal. Thoughtfully designed self-evaluation procedures that provide students with explicit criteria at an appropriate level of generality, that provide for student involvement in assessment decision making, that elicit student cognitions about their performance, which ground student goal setting in accurate data,

and that are integrated with sensitive instruction may provide teachers with a powerful lever for enhancing student learning.

References

- Arter, J., Spandel, V., Culham, R., & Pollard, J. (1994, April). The impact of training students to be self-assessors of writing. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.
- Bandura, A. (1997). Self-efficacy: The exercise of control. New York: W. H. Freeman.
- Boud, D. J., Churches, A. E., & Smith, E. M. (1986). Student self-assessment in an engineering design course: An evaluation. International Journal of Applied Engineering Education, 2(2), 83-90.
- Butler, R. (1990). The effects of mastery and competitive conditions on self-assessment at different ages. Child Development, 61, 201-210.
- Chrispeels, J. H. (1997). Educational policy implementation in a shifting political climate: The California experience. American Educational Research Journal, 34(3), 453-481.
- Crocker, A. C., & Cheeseman, R. G. (1988). The ability of young children to rank themselves for academic ability. Educational Studies, 14(1), 105-110.
- Dorfman, A. (1997, April). Teachers' understanding of performance assessment. Paper presented at the annual meeting of the American Educational Research Association, Chicago.
- Falchikov, N., & Boud, D. (1989). Student self-assessment in higher education: A meta-analysis. Review of Educational Research, 59(4), 395-430.
- Fuchs, L. S., Fuchs, D., Karns, K., Hamlett, C., Katzaroff, M., & Dutka, S. (1997). Effects of task-focused goals on low-achieving students with and without learning disabilities. American Educational Research Journal, 34(3), 513-543.
- Gibson, S., & Dembo, M. H. (1984). Teacher efficacy: A construct validation. Journal of

Educational Psychology, 76(4), 569-582.

Glass, G., McGaw, B., & Smith, M. (1981). Meta-analysis in social research. Beverly Hills, CA: Sage.

Hillocks, G. (1986). Research on written composition: New directions for teaching. Urbana, IL: ERIC Clearinghouse on Reading and Communication Skills.

Khattari, N., Kane, M., & Reeve, A. (1995). How performance assessments affect teaching and learning. Educational Leadership, 53(3), 80-83.

Koretz, D., Stecher, B., Klein, S., & McCaffrey, D. (1994). The Vermont portfolio assessment program: Findings and implications. Educational Measurement: Issues and Practice, 13(3), 5-16.

Kulm, G. (1994). Mathematics assessment: What works in the classroom. San Francisco: Jossey-Bass.

Lee, S., & Suen, H. (1995, April). Effect of two assessment techniques on study strategy use. Paper presented at the annual conference of the American Educational Research Association, San Francisco.

Linn, R. (1997). Evaluating the validity of assessments: The consequences of use. Educational Measurement: Issues and Practice, 16(2), 14-16.

Lipson, M. Y., & Mosenthal, J. (1997, March). The differential impact of Vermont's writing portfolio assessment on classroom instruction. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Locke, E., Shaw, K., Saari, L., & Latham, G. (1981). Goal setting and task performance: 1969-1980. Psychological Bulletin, 90(1), 125-152.

Markman, E. (1979). Realizing that you don't understand: Elementary school children's awareness of inconsistencies. Child Development, 50, 643-655.

Meece, J., Blumenfeld, P., & Hoyle, R. (1988). Students' goal orientations and cognitive engagement in classroom activities. Journal of Educational Psychology, 80(4), 514-523.

Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. Educational Measurement: Issues and Practice, 14(4), 5-8.

Moss, P. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. Review of Educational Research, 62(3), 229-258.

Pajares, F. (1996). Self-efficacy beliefs in academic settings. Review of Educational Research, 66(4), 543-578.

Paris, S., Lawton, T., Turner, J., & Roth, J. (1991). A developmental perspective on standardized achievement testing. Educational Researcher, 20(5), 12-20, 40.

Paris, S., Turner, J., & Lawton, T. (1990). Students' views of standardized achievement tests. Paper presented at the annual meeting of the American Educational Research Association, Boston.

Paris, S. G., Turner, J. C., Muchmore, J., & Perry, N. (1995). Teachers' and students' perceptions of portfolios. Journal of Cognitive Education, 5(1), 6-40.

Rhodes, L. K. (Ed.). (1993). Literacy assessment: A handbook of instruments. Portsmouth, NH: Heinemann.

Rolheiser, C. (Ed.). (1996). Self-evaluation...Helping kids get better at it: A teacher's resource book. Toronto: OISE/UT.

- Ross, J. A. (1995b). Strategies for enhancing teachers' beliefs in their effectiveness: Research on a school improvement hypothesis. Teachers College Record, 97(2), 227-251.
- Ross, J. A. (1998). The antecedents and consequences of teacher efficacy. In J. Brophy (Ed.) Advances in Research on Teaching. Vol. 7. (pp. 49-74) Greenwich, CT: JAI Press
- Ross, J. A., Rolheiser, C., & Hogaboam-Gray, A. (1998a). Skills training versus action research in-service: Impact on student attitudes to self-evaluation. Teaching and Teacher Education, 14(5), 463-477.
- Ross, J. A., Rolheiser, C., & Hogaboam-Gray, A. (1998b). Student evaluation in cooperative learning: Teacher cognitions. Teachers and Teaching, 4(2), 299-316.
- Ross, J. A., Rolheiser, C., & Hogaboam-Gray, A. (1999). Teaching students how to self-evaluate when learning cooperatively: The effects of collaborative action research on teacher practice. Elementary School Journal, 99(3), 255-274.
- Schunk, D. H. (1996, April). Self-efficacy for learning and performance. Paper presented at the annual meeting of the American Education Research Association, New York.
- Schunk, D. H. (1995, April). Goal and self-evaluative influences during children's mathematical skill acquisition. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.
- Serlin, R. (1993). Confidence intervals and the scientific method: A case for Holm on the range. Journal of Experimental Education, 61(4), 350-360.
- Shepard, L. (1997). The centrality of test use and consequences for test validity. Educational Measurement: Issues and Practice, 16(2), 5-8,13,24.
- Shepard, L. A., Flexer, R. J., Hiebert, E. H., Marion, S. F., Mayfield, V., & Weston, T. J.

(1996). Effects of introducing classroom performance assessments on student learning.

Educational Measurement: Issues and Practice, 15(3), 7-18.

Slater, T., Ryan, J., & Samson, S. (1997). Impact and dynamics of portfolio assessment and traditional assessment in a college physics course. Journal of Research in Science Teaching, 34(3), 255-271.

Stiggins, R. (1994). Student-centered classroom assessment. New York: Merrill.

Stipek, D., Recchia, S., & McClintic, S. (1992). Self-evaluation in young children. Monographs of the Society for Research in Child Development, 57, 1-84.

Stipek, D., Roberts, T., & Sanborn, M. (1984). Preschool-age children's performance expectations for themselves and another child as a function of the incentive value of success and the salience of past performance. Child Development, 59, 1983-1989.

Urdan, T., & Maehr, M. (1995). Beyond a two-goal theory of motivation and achievement: A case for social goals. Review of Educational Research, 65(3), 213-243.

van Kraayenoord, C., & Paris, S. (1993, December). Self-assessment: Children's perspectives on their literacy activities in the classroom. Paper presented at the National Reading Conference, Charleston, SC.

Wagner, M. J. (1991). Goal-setting, self-efficacy, and learning performance: A causal model. Unpublished doctoral dissertation, University of Arizona, AAC 911978 ProQuest - Dissertation Abstract.

Wiggins, G. (1993). Assessing student performance: Explore the purpose and limits of testing. San Francisco, CA: Jossey-Bass.

Table 1: Means, Standard Deviations, and Reliabilities of Teacher Variables

	Number of Items	Range	Alpha	Treatment (N=15)		Control (N=15)	
				Mean	<u>SD</u>	Mean	<u>SD</u>
Assessment Practices	10	1-6	.84	3.88	.66	3.79	.97
Personal Teaching Efficacy	9	1-6	.69	4.18	.49	4.30	.50
General Teaching Efficacy	6	1-6	.77	3.99	.64	4.03	1.02
Mean years of teaching				13.40		16.07	
Gender							
male				5		5	
female				10		10	
Grade of class							
3/4				1		0	
4				3		4	
4/5				0		2	
5				5		3	
5/6				3		2	
6				3		4	
Academic Training							
BA/BSc				13		13	
AQ Courses				8		7	
Principal Course				1		1	
MA/Med				1		1	
Conferences				11		9	
Summer Institutes				6		6	

Table 2: Internal Reliability of Student Variables (N=290)

	Number of Items	Range	Cronbach's Alpha
Self Efficacy	6	1-10	.85
Self-Evaluation	6	1-10	.84
Self-Evaluation Attitudes	10	1-5	.75
Goal Orientations			
mastery	9	1-5	.84
ego	3	1-5	.62
affiliative	3	1-5	.54

Table 3: Unadjusted Means and Standard Deviations of Student Variables, by Experimental Condition

	Treatment (N=148)		Control (N=148)	
	<u>M</u>	<u>SD</u>	<u>M</u>	<u>SD</u>
Self-Efficacy	7.40	1.58	7.28	1.88
Self-Evaluation	7.34	1.69	7.03	1.66
Self-Evaluation Attitudes	3.85	.63	3.97	.62
Goal Orientations				
mastery	3.77	.75	3.88	.76
ego	2.92	.96	2.98	.95
affiliative	3.21	1.04	3.17	1.16
Achievement				
pre	3.62	1.37	2.90	1.22
post	4.17	1.23	3.23	1.31
	Percent		Percent	
Accurate				
pre	35		26	
post	64		40	
Gender				
male	54		51	
female	46		49	
Age				
9 or under	23		23	
10	37		40	
11	36		29	
12 & over	5		8	

Table 4: Posttest Accuracy of Underestimating, Accurate, and Overestimating Students, by Experimental Condition

Pretest Accuracy	% Accurate on Posttest	Statistical Significance
Underestimate (N=5) treatment (N=4) control (N=1)	50 0	Fisher's exact test $p=.600$
Accurate (N=90) treatment (N=53) control (N=37)	72 54	$\chi^2(1, 90)=2.960$, $p=.085$
Overestimate (N=202) treatment (N=91) control (N=111)	30 18	$\chi^2(1, 02)=3.803$, $p=.051$.

Table 5: Effect of Self-Evaluation on Student Achievement: Results of Analysis of Covariance

Source of Variation	SS	DF	MS	F	p	Partial Eta Sq _d
Within + residual	3.35	291	.01			
Pretest	1.30	1	1.30	112.98	<.001	.280
Self-Evaluation Accuracy	.01	1	.01	1.30	.255	.004
Treatment	.08	1	.08	7.22	.008	.024
Pretest x Treatment	.11	1	.11	9.50	.002	.032
Self-Evaluation x Treatment	.00	1	.00	.03	.859	.000
Model	2.50	5	.50	43.44	<.001	
Total	5.86	296	.02			
R-squared	.427					

Table 6: Adjusted Posttest Achievement Means and Standard Deviations of Low and High Achieving Groups, by Experimental Conditions

	Low Pretest Achievement		High Pretest Achievement	
	Treatment (<u>N</u> =73)	Control (<u>N</u> =109)	Treatment (<u>N</u> =75)	Control (<u>N</u> =40)
Pretest Mean	.53	.51	.76	.73
Pretest SD	.12	.12	.01	.01
Posttest Mean	.65	.56	.75	.71
Posttest SD	.11	.14	.01	.11

Appendix: Assessment Scales for Junior Division Narrative

Level 6

The work demonstrates a confident command and integration of all the elements of writing. The content is often strikingly creative and imaginative (e.g., evidence of risk taking).

Possible Characteristics

- The controlling idea and its development are insightful and original, and consistent with the narrative form.
- The organization is subtle; the control is secure; the style reinforces the purpose; events are well sequenced with supporting detail.
- The voice is confident; there is a sense of engagement with the topic and an effective relationship with the audience; voice is appropriate to the narrative words.
- The control of written conventions of language is skilful; rare errors in spelling and minor errors in grammar and punctuation may exist but do not affect the overall impact; they may be the result of the difficulty of the writing task and/or risks taken by the student.

Level 5+*Possible Characteristics*

- Writer is in command of elements of narrative but not completely.
- Controlling idea original and creative but not striking.
- Development of ideas demonstrate originality.
- Voice is clear and effective.
- Strong sense of reader.
- Effective level of word choice.
- Very few errors in conventions; spelling errors a result of use of difficult word choice.

Level 5

The work shows an effective control and integration of all the elements of writing. The content is thoughtful and thorough.

Possible Characteristics

- The controlling idea and its development are thoughtful and thorough, and consistent with the narrative form.
- The organization is effective; the style is appropriate to the purpose and the narrative form.
- The voice is clear; there is a strong sense of audience.
- The control of the written conventions is sound; any errors in spelling, grammar, and punctuation do not detract from the overall impact.

Level 4+*Possible Characteristics*

- Good control of the elements of narrative writing.
- Some evidence of originality.
- Consistent narrative voice; good awareness of audience.
- Organization is clearly evident.
- Good evidence of style.
- Conventions in good control; there may be some errors but they do not detract from meaning.

Level 4

The work shows control of the elements of writing. It is generally integrated. The content is clear and complete.

Possible Characteristics

- The controlling idea and its development are clear but may be conventional or derivative (e.g., a summary of events).
- Organization is capable; there is a clear attempt to connect style and purpose with narrative form.
- The voice is apparent but may fluctuate; there is an awareness of audience.

The control of the written conventions is capable; infrequent errors may detract from the overall impact of the work but do not affect the meaning.

Level 3+*Possible Characteristics*

- Writer makes an obvious effort to involve the reader.
- Elements of narrative writing are under control but any lack of control can effect meaning.
- Integration of elements, development of story is almost complete.
- Organization is apparent but still not really clear; paragraphing is used.
- Evidence of a narrative voice, but not consistent.
- Control of conventions is capable but still can impact on meaning.

Level 3

The work shows control of most of the elements of narrative writing. Some integration is apparent. The content may be simple or unoriginal.

Possible characteristics

- The controlling idea and its development are apparent and show some balance or consistency; ideas convey surface meaning.
- Organization is apparent; there is some attempt to connect style and purpose.

- There is a sense of voice with some control; there is an occasional awareness of audience.
- Control of the written conventions of language is evident; errors occasionally detract from the impact and the meaning.

Level 2+*Possible Characteristics*

- A firm grasp of the basic elements (conventions, sentence structure—not necessarily paragraphs).
- Limited sophistication/maturity of ideas.
- Controlling idea is apparent but uneven.
- Some organization is apparent, but little or no attempt to connect style to purpose/theme.
- Narrative voice emerging but distinction between writer's personal voice and the narrative voice not clear (writer-oriented text vs. reader-oriented text).
- Conventions distract but understanding of ideas is possible.

Level 2

The work shows grasp of some of the basic elements of narrative writing; the writing conveys simple ideas.

Possible Characteristics

- The controlling idea and its development are limited but discernible; ideas are superficial.
- Organization is attempted; style is simple and unconnected to the purpose.
- Voice may be often limited to a personal, vernacular register; awareness of audience is limited or absent.
- Grasp of the written conventions of language is tentative; errors are distracting and often interfere with the reader's understanding of the ideas.

Level 1

The work shows a minimal grasp of some of the basic elements of writing. The content conveys unconnected or fragmented ideas.

Possible Characteristics

- The writing expresses some unconnected ideas, but no discernible controlling idea.
- Organization is not discernible.
- Voice is limited to personal, vernacular register; awareness of audience is absent.
- Grasp of the written conventions of language is minimal; errors impede expression and comprehension.

Additional Scoring Notes

If the work was less than half a page, the passage was scored no higher than

Endnotes

¹ The research was funded by the Social Sciences and Humanities Research Council of Council, the Ontario Ministry of Education, and Durham Region District Board. The views expressed in the report do not necessarily reflect the views of the Council, the Ministry or the school district.

² Although the pretest differences could have been just bad luck, it is possible that teachers might have been unwilling to risk a change in teaching practice with a less able class. If so, these teachers might not have volunteered for the treatment condition, skewing the ability distribution. Control group teachers with less able classes might have felt no such constraint because no change in their practice was required. It is also possible that a Hawthorne effect might have inflated pretest scores in the treatment condition.

³ We subtracted the pretest differences between treatment and control means from the posttest treatment-control mean differences and divided by the pretest control group standard deviation (Glass, McGaw, & Smith, 1981).

⁴ Both treatment and control groups were given the criteria used to mark their essays but neither received the level descriptions in the Appendix for several reasons: (a) We thought it would be in conflict with the treatment to impose a rubric on students without inviting them to participate in its construction. (b) Imposing a rubric would not be a problem for the control students but giving only them the rubric would disadvantage the treatment group. (c) We did not want to involve treatment students in the construction of the posttest rubric because if only they had the rubric it would disadvantage the control. (d) We did not want to involve both groups in rubric construction because that would dilute the differences between the conditions.