Student Self-Evaluation, Teacher Evaluation, and Learner Performance

Zane Olina
Howard J. Sullivan

A total of 341 Latvian students and eight teachers participated in this study of student self-evaluation and teacher evaluation. Students completed a 12-lesson teacher-directed instructional program on conducting and writing a report of their own experimental research. Sixteen classes were randomly assigned to one of four treatment conditions: (1) no in-program evaluation, (2) self-evaluation and revision at the research design and draft final report stages, (3) teacher evaluation and student revision at both stages, (4) self-plus-teacher evaluation and student revision at both stages. Students in the teacher-evaluation and self-plus-teacher evaluation conditions received significantly higher ratings from an independent rater on their final research reports. However, students under the self-evaluation conditions had greater confidence in their ability to conduct future experiments.

In recent years several authors have called for classroom assessment practices that are closely integrated with instruction, used formatively to support student learning, and involve students actively in evaluating their own work (Bransford, Brown & Cocking, 1999; Gipps, 1994; Shepard, 2000; Wiggins, 1998; Wolf, Bixby, Glenn, & Gardner, 1991). One way of addressing these goals is through use of formative classroom assessment. Sadler (1989) contended that formative assessment is concerned with how judgments about the quality of student responses can be used to improve student learning. He distinguished between two formative assessment strategies according to the source of evaluative information—(a) self-monitoring or self-evaluation where the relevant information is generated by the learner and (b) feedback where the information source is external to the learner (Sadler). The current study examined the effects of these two formative assessment strategies, student self-evaluation and teacher evaluation, the most common source of external feedback, on student learning.

The rationale for developing student self-evaluation skills comes from several areas of research, including self-regulated learning (Ertmer & Newby 1996; Schunk, 1989; Zimmerman, 1989), metacognition (Bransford et al., 1999; Flavell, 1976; Winne & Hadwin, 1998), and classroom assessment (Gipps, 1994; Sadler, 1989; Shepard, 2000; Stiggins, 2001). Researchers on self-regulated learning view self-evaluation as an effective self-regulated learning strategy. Other such strategies include seeking, organizing, and transforming information; seeking social assistance; setting goals; and planning (Zimmerman). In addition, positive self-evalua-

tion is thought to increase student self-efficacy beliefs (Bandura, 1977). Students with greater self-efficacy beliefs may set higher personal performance standards.

Bransford et al. (1999) viewed self-evaluation as part of a metacognitive approach to instruction that develops "people's abilities to predict their performances on various tasks and to monitor their current levels of mastery and understanding" (Bransford et al., p. 12). Teaching practices that use a metacognitive approach to learning focus on self-assessment and reflection on what worked and what needs improving. Within this framework, self-evaluation can help learners appraise their current understanding in order to determine improvement needs. Most research on metacognition has focused on developing student ability to monitor their learning behaviors through goal-setting, record-keeping, using job aids or cuing devices to check for understanding, and other strategies (Bransford et al.).

Some classroom assessment researchers emphasize the need for assessment to become an integral part of instruction (Gipps, 1994; Shepard, 2000; Stiggins, 2001). The supporters of this approach see student self-evaluation as a means for making instruction more interactive, involving students more actively in their own learning, and developing lifelong learning skills (Black & William, 1998b; Sadler, 1989; Wolf et al., 1991). Classroom assessment researchers view self-evaluation as an instructional strategy that, when applied to the mastery of specific learning tasks, can help students better understand the learning goals and take greater responsibility for their own learning. Sadler went even further, suggesting that part of teachers' responsibility is "to download [their] evaluative knowledge so that students eventually become independent of the teacher and intelligently engage and monitor their own development" (p.141).

Even though the three aforementioned fields of self-regulation, metacognition, and classroom assessment study different aspects of self-evaluation, they share a similar definition of self-evaluation and regard it as a worthwhile learning skill enabling students to learn more effectively and improve their performance. Researchers on self-regulated learning describe self-evaluation as a characteristic of effective learners. From this perspective, the question of how one becomes a self-regulated learner needs further exploration. Researchers in metacognition are interested in self-evaluation as a fine-grained cognitive process of monitoring one's understanding. For classroom assessment researchers, self-evaluation is an instructional strategy for engaging students actively in learning.

Overall, studies show positive effects of self-evaluation on student performance and motivation across subject areas and age groups. Klenowski (1995) conducted a qualitative study on the use of self-evaluation processes of secondary students in England and Australia. Both teachers and students in this study found self-evaluation valuable and reported that it helped increase student confidence and that students became more self-critical as a result. Maqsud and Pillai (1991) found that South African high school students who were asked to self-score their tests over a course of one semester significantly outperformed students whose tests were scored by the teacher. Several authors reported that primary school students who received self-evaluation training over a period of several weeks in a specific subject area performed better in mathematics (Fontana & Fernandez, 1994) and narrative writing (Ross, Rolheiser, & Hogaboam-Gray, 1999) than their counterparts who were not trained in self-evaluation.

Authors of the studies discussed above suggest several reasons for the positive effects of self-evaluation on student performance, such as clarifying learning objectives for the students, reinforcing their previous learning, and providing teachers with additional information about student learning that teachers can then use to improve their instruction (Fontana & Fernandez, 1994; Ross et al., 1999). A different explanation for the effects of self-evaluation comes from theories of motivation and self-regulation. Maqsud and Pillai (1991) suggested that most likely the self-scoring group in their study scored higher on the final test because they had attributed their lower results on earlier assessments to lack of effort rather than to factors beyond their control, such as ability, task difficulty, luck, poor construction of test, or subjective evaluation by the teacher. When students

think that their failure is due to poor effort, they may study harder the next time (Schunk, 1989).

Studies about the effects of teacher evaluation on student performance and attitudes have yielded somewhat mixed results. Cardelle-Elawar and Corno (1985) found that elementary school student performance and attitudes toward mathematics improved when teachers provided written feedback on their homework several times a week. Thomas et al. (1993) reported a positive correlation between the amount of teacher feedback on tests, quizzes and homework assignments and student performance in high school biology courses. Page (1958) found in his study involving 74 secondary school teachers that a brief written comment on objective examinations significantly improved student performance when compared to no comment at all. Cardelle-Elawar and Corno suggested that the main reason teacher evaluation results in improved student performance is that specific feedback on errors draws student attention to material not adequately learned and provides corrective guidance.

Other studies have shown no effect of teacher evaluation on student performance. Stewart and White (1976) replicated Page's (1958) study and reviewed 12 other replication studies, concluding that teacher comments had little or no effect on student performance. In their review of more than 250 studies of formative assessment, Black and William (1998a) concluded that teacher evaluation can result in positive and negative effects on student learning depending on the quality of feedback. For example, Butler (1988) reported that when 11-year-old Israeli students were given three types of written teacher feedback, their interest and performance was highest after receiving comments only, whereas receiving grades only or grades plus comments had similar and generally undermining effects on their perfomance.

Studies involving comparisons of teacher evaluation and student self-evaluation have shown positive effects of self-evaluation on student continuing motivation. Salili, Maehr, Sorensen, and Fyans (1976) found no performance differences among fifth-grade Iranian students who solved word anagram problems under three different evaluation conditions. However, they reported that students in the self-evaluation and peer-comparison conditions showed greater continuing motivation than those in the teacher evaluation condition in the form of desire to return to initial task. Hughes, Sullivan, and Mosley (1985) reported that students returned to a difficult task more often after self-evaluation and to an easy task more often after teacher evaluation. Salili et al. suggested that their study confirmed the hypothesis that the addition of extrinsic rewards, that is, teacher evaluation, reduces intrinsic interest. Thus, students in the teacher-evaluation condition were less likely to return to the initial task. Hughes et al. (1985) contended that the two conditions—(a) a hard task on which they are likely to perform more poorly than on an easy task and (b) knowledge that their performance will be observed by an external evaluator, that is, a teacher—pose a threat that reduces student motivation to return to a task.

The inconclusive findings about the effects of self-evaluation and teacher evaluation and the considerable role that student attitudes seem to play in their interpretation and subsequent use of evaluative feedback prompted the authors to conduct this study. Four levels of evaluation were compared: (a) a self-evaluation condition, (b) a teacher-evaluation condition, (c) a self-plus-teacher evaluation condition, and (d) a no-evaluation control condition. The teacher evaluation component in this study was designed to incorporate several strategies identified through a review of evaluation research. The review indicated that (a) effective teacher evaluation should be specific, and directly related to the task; (b) evaluation criteria should be made explicit to the students before they begin working on the learning task, and (c) students should be provided with opportunities to revise their work after it is evaluated (Black & William, 1998a; Hughes et al., 1985; Sadler, 1989; Wiggins, 1998). Thus, we provided students with specific standards for evaluating their work before they began working on their experiments. These standards were first made explicit to the students in the form of a Project Rating Scale constructed for the study. The students under the teacher-evaluation condition received teacher comments about their work according to

criteria in the Project Rating Scale. The students were not assigned grades on their initial reports, and they had opportunities to revise their work on their final reports.

The student self-evaluation component was designed to be similar to the teacher-evaluation component. Students were provided with the same specific standards for evaluating their own work in the form of the Project Rating Scale that the teachers received. The students then applied the scale to evaluate their written reports and to write comments about their work. The students were not assigned grades on their initial reports, and they had opportunities to revise their work on their final reports. This approach was consistent with strategies for self-evaluation training suggested by several authors (Rolheiser, 1996; Sadler, 1989; Wiggins, 1998).

Consistent with prior research, the following core questions were investigated:

1. What are the comparative effects of teacher evaluation and student self-evaluation on student performance?
2. Does the combination of teacher evaluation and student self-evaluation improve student performance to a greater degree than either of the evaluation procedures alone?
3. Do the two evaluation procedures have differential effects on student attitudes?

As an extension of prior research, we also examined data related to other issues. These included overall student and teacher perceptions of self-evaluation and teacher evaluation, the manner in which the two particular strategies affected student work, and difficulties students and teachers might experience when attempting to use the two formative evaluation strategies in a regular classroom setting. Further, several areas of the study contribute to its uniqueness. First, we used a complex performance-based task as the main criterion measure, whereas learning outcomes used in most prior research conducted on self- and teacher evaluation can be assessed by quizzes and progress tests that can be scored correct or incorrect. Second, the study was conducted in a naturalistic classroom setting and implemented by teachers in their regular classrooms over an extended period of time. Third, the study was conducted in Latvia, home country of the first author, where the authoritarian classroom management style of teachers still dominates in most classrooms, and there is no tradition of engaging students in evaluation of their own work. Since Latvian independence from Russia in 1990, the educational reform movement in Latvia has emphasized the need to develop lifelong learning skills, including the ability to evaluate one's own work.

This study was based in part on the findings of earlier research (Olina & Sullivan, 2002) that investigated the effects of teacher evaluation and the combination of teacher evaluation and student self-evaluation on the performance and attitudes of 189 Latvian high school students. In our earlier study, students in the teacher-evaluation and the self-plus-teacher evaluation groups received significantly higher ratings on their final projects than those in the no-evaluation group. However, the no-evaluation group had more favorable attitudes toward the program than the other two groups, while the self-plus-teacher evaluation group was significantly more confident of their ability to independently conduct future experiments. These results indicating the beneficial effect of teacher evaluation on student performance and potential positive effect of self-evaluation on student attitudes prompted us to conduct another study in order to better understand the unique contribution of self-evaluation to student learning and to explore the relationship between teacher and self-evaluation.

Specifically, we extended the earlier research in several ways. A self-evaluation-only condition, which was not included in the previous study, was added to the design. Moreover, the students in all treatment groups were trained on the use of the Project Rating Scale, a rubric that would be used for evaluating student research projects, whereas only the self-plus-teacher evaluation group was trained in the previous study. This change enabled us to better isolate the effects of formal engagement in self-evaluation. Results in the previous study could be attributed partially to the fact that the self-plus-teacher evaluation group was trained on the use of the Project Ratings Scale and, thus, knew the project evaluation criteria better than did the other two

groups. The current study was also broader in scope than the earlier study, and involved a greater number of teachers and students. A four-hour training workshop for teachers was added to better prepare them for the instruction, as well as several additional data sources, such as student focus groups, analysis of student draft reports, and student and teacher evaluations of these reports, to gain more in-depth information about program delivery and participant attitudes.

Based on the results of our initial study and review of literature, we anticipated that the teacher evaluation would improve student performance to a greater extent than would student self-evaluation. Teachers are knowledgeable about conducting experiments and writing reports and thus are likely to give students more accurate feedback on student performance and corrective guidance than can the students themselves. We predicted that students in the self-evaluation group would perform better than students in the no-evaluation condition, because use of the Project Rating Scale might clarify the final task requirements for them and they might be able to focus on what they needed to improve. In addition, engagement in evaluating their own work against a set of objective criteria is likely to reduce the influence of other factors that students could use to explain lower ratings, such as subjective evaluation by the teacher, or luck. As a result, students might put more effort into improving their work.

METHOD

Subjects

Participants were 341 10th- and 11th-grade Latvian students from 16 classes taught by eight different teachers. The average class size was 21 students. Of the classes, 12 were 10th-grade and 4 were 11th-grade. All teachers involved in the study had completed at least four years of college and obtained either a bachelor's or a master's degree in education.

The classes were drawn from eight schools in different regions of Latvia, representing both rural and urban areas, and varied socio-economic backgrounds. One school was located in the capital, Riga, with approximately 800,000 inhabitants; three schools were in cities of 30,000 to 90,000; three schools were in towns of approximately 10,000; and one was in a suburban community of approximately 1,500. Larger urban centers in Latvia tend to attract students from a larger geographic area, are more selective in terms of their student body, and attract higher-quality teachers. All schools in this study had somewhat similar high-school entrance requirements. As part of admission requirements, the students in all schools had to meet or exceed the same minimum grade point average. The researchers did not anticipate that the location of the school or socio-economic status of individual students would significantly affect student achievement in this study, because all schools in Latvia are expected to meet the same educational standards in each subject area, as mandated by the Latvian Ministry of Education and Science. Moreover, there is little tradition of teaching process skills, such as designing experiments and writing research reports, to the students and, thus, concepts included in the instructional program in this study would be new to most teachers and students.

Materials

A 12-lesson instructional program entitled *Learning Explorations* was developed in print form in Latvian for use in the study. The program was designed to teach students about basic experimental design concepts (such as hypotheses, dependent variables, control groups, independent variables, treatment conditions, and constants) and about common design flaws. Students were also instructed on how to write simple experiment reports. Upon completion of this instruction, students designed and conducted their own experiments about learning, in response to a scenario calling for them to test whether presentation of information in an organized way facilitated learning. They then produced a simple written report of the results. Before they began work on their projects, students were introduced to the Project Rating

Scale (described below) that they could use for self-evaluation. Students in all treatment groups practiced applying the scale to an example of both a poor and an excellent student report. The program was intended for use in introductory psychology classes, or as supplemental material in other classes in which students were introduced to the design of independent research projects.

The *Learning Explorations* program was organized into six sections, each of which required about two 40-min class periods. In addition to explanations of key concepts and examples, students were provided with multiple practice opportunities through analysis of brief experiment scenarios related to human learning and engagement in a variety of interactive teacher-directed activities. The program consisted of a student book and a teacher guide. The student book contained all the information presented during instruction, practice exercises, worksheets, and examples of experiments. The teacher guide included step-by-step lesson plans on how to use the student book, descriptions of instructional activities, a posttest, the Project Rating Scale for rating student projects, transparencies, and handout masters. All materials used in the study were in the Latvian language. The instructional program and the assessment instruments were field tested in the earlier research (Olina & Sullivan, 2002) with approximately 190 students from a similar age group to the present population.

Criterion Measures

Four criterion measures were used in the study: (a) ratings of the student projects, (b) posttest scores, (c) student attitude surveys, and (d) teacher attitude surveys.

Ratings of student research projects. Student project rating scores served as the primary criterion for assessing student performance. Student projects were evaluated based on the Project Rating Scale, a descriptive scale consisting of 15 items developed specifically for this study. Each item was rated on a 3-point scale from 2 (*above average*) to 0 (*below average*). The maximum number of points on the rating scale was 30. The 15 criteria included in the Project Rating Scale are listed in Table 1. A sample item from the Project Rating Scale is provided in Table 2. The description in the left column was assigned 2 points, that in the middle column, 1, and for the description in the right column, 0, as shown in the sample item.

Table 1 ☐ Criteria on the Project Rating Scale.

Introduction
1. Describes the independent variable.
2. Describes the dependent variable.

Method
3. The hypothesis is clear and specific.
4. Treatment conditions are fully described for all groups.
5. Constants are similar for all groups.
6. At least five participants are in each group.
7. Describes experimental procedures step-by-step.
8. Provides clear instructions for the participants.
9. Participants are randomly assigned to groups.

Results
10. Compares mean scores for groups.
11. Evaluates the hypothesis based on the data.

Conclusions
12. Provides a feasible explanation of the results.
13. Provides suggestions for potential application of the results.

Overall Effort
14. The work overall is thoughtful and shows good understanding of experimental design concepts.
15. The work overall is accurate.

The first author of this paper rated all student projects. Another trained rater evaluated 10% of all student projects to determine interrater reliability. The other rater was a former classroom teacher with considerable experience rating student projects. The raters were blind to the student's name, experimental condition, and school of origin for each project. Internal reliability of the Project Rating Scale, using Cronbach's alpha, was .87. The Pearson correlation coefficient for the interrater reliability between the ratings of the final student projects by the first author and the independent rater was .90. The first author's ratings were used for the analysis.

Posttest. The posttest served as a second criterion measure for assessing student performance.

Table 2 ☐ Sample item from the Project Rating Scale.

Item: Describes experimental procedures step-by-step.		
❒ describes experimental procedures step-by-step, including all materials and conditions	❒ some steps and conditions are missing, but overall the experiment can be carried out by following the procedures	❒ most steps are missing, no materials and conditions are included

The posttest items were directly aligned with the objectives of the instructional program. They required recognition and recall of concepts and application of knowledge to scenarios provided by the experimenter. The posttest required less depth of understanding of the program content than designing and conducting the research projects.

The posttest consisted of 21 multiple-choice and short-answer items, and had a maximum score of 30 because some items had multiple-point answers. Two posttest items requiring students to write a summary of results of a given experiment scenario were worth 3 points each, and one item requiring students to write an experimental procedure for a given scenario was worth 6 points. These three items were scored using a descriptive rating scale developed for that purpose. Internal reliability of the posttest, using Cronbach's alpha, was .75. A sample multiple-choice item from the posttest is provided below.

Which of the following is the factor that is changed on purpose? Circle the correct answer.

a. Constant.

b. Control group.

c. Dependent variable.

d. Independent variable.*

* Correct answer

Student Attitude Survey. A core six-item attitude survey, with added items for subjects in the self-evaluation and the teacher-evaluation conditions, served as the criterion measure for assessing students' attitudes and motivation toward the instruction. The survey consisted of Likert-type questions with four response choices ranging from *strongly agree* to *strongly disagree.* The items dealt with topics such as whether students liked the program, whether they liked conducting experiments, and whether they were confident in their ability to conduct experiments and write research reports as a result of the program. Two additional Likert-type items about self-evaluation were added to the six-item student survey for subjects in the self-evaluation group and two additional items about teacher evaluation were added for subjects in the teacher-evaluation condition. These items asked students whether they liked self-evaluation or teacher evaluation respectively and whether self-evaluation or teacher evaluation helped them improve their work. All four of these additional items were added to the survey for the self-plus-teacher evaluation condition.

Teacher Attitude Survey. A core 11-item attitude survey served as the criterion measure for assessing teacher attitudes. Of the items, 8 were four-choice Likert-type questions with response options ranging from *strongly agree* to *strongly disagree.* These items assessed teacher attitudes toward the appropriateness of the instruction for the students, student progress, ease of program delivery, and teacher satisfaction with their own work as instructors. Three remaining items asked teachers to report on the program implementation process and to make suggestions for improvement of the instructional program. As with the student attitude survey, additional Likert-type items about self-evaluation and teacher evaluation were added as appropriate to the teacher survey for the self-evaluation, teacher-evaluation and self-plus-teacher evaluation conditions.

Responses to the Likert-type items in both the

student and the teacher surveys were assigned a score of 3 (*strongly agree*) for the most positive response and a score of 0 (*strongly disagree*) for the least positive response.

Procedures

Participant assignment to treatments. The experimenter used a systematic procedure to assign the eight schools to one of four treatment conditions (no evaluation, self-evaluation, teacher-evaluation, and self-plus-teacher evaluation) by first blocking the eight schools into two groups by ability (four lower- and four higher-ability schools) and then randomly assigning schools from each block to one of the four treatments. The composite score on the ninth-grade standardized mathematics examination and the Latvian language examination served as the achievement measure for this purpose. These exams are prepared by the Latvian Ministry of Education and Science and are administered in a centralized manner across the country. It is the only standardized measure of achievement that exists for the population involved in the study.

The composite ability score on the two exams for each student could range from 2 to 20 because each exam is scored on a 1-to-10 point scale, with a score of 1 as the lowest and a score of 10 as the highest. The mean score was 11.27 (SD = 2.58) for the lowest-ability school and 13.51 (SD = 2.40) for the highest-ability school, with the overall mean ability score for all schools in the study of 12.26 (SD = 2.58). Upon assignment to treatments, student mean ability scores were as follows: 11.81 (SD = 2.28) for the no-evaluation group, 12.38 (SD = 2.40) for the self-evaluation group, 12.11 (SD = 2.98) for the teacher-evaluation group, and 12.68 (SD = 2.63) for the self-plus-teacher evaluation group. A one-way analysis of variance (ANOVA) yielded no significant differences between groups on the ability measure. In addition, the treatment groups were quite similar in terms of student socioeconomic backgrounds and types of communities that the students represented.

Teacher expectations and training. All teachers received the same version of the instructional program. Teachers in the no-evaluation group received no additional instructions for use of the program. Teachers in the remaining three treatments received additional instructions describing the evaluation procedures that they were expected to complete for their treatment condition.

The teachers were told that the purpose of the research study was to investigate the effects of self-evaluation and teacher evaluation on student performance. They were told that there were four treatment groups in the study and that each group would be doing slightly different things in the program as specified in the instructions.

The teachers participated in a 4-hr training workshop one week prior to the start of the instruction. The main purpose of the workshop was to ensure effective and uniform delivery of the basic instructional program. During the training workshop, the teachers were introduced to the overall sequence of the basic instructional program, learned how to conduct one of the demonstration experiments included in the program, examined common errors that students made in their reports, and practiced scoring sample student research projects using the Project Rating Scale. Teachers also learned the additional procedures pertaining to their treatment condition in small groups by treatment.

Student expectations from the program. To ensure student motivation throughout the study, all students received a grade for their research projects and posttest. They were also told that all students who completed the experiment report, the posttest, and the student survey would participate in a lottery for a valuable prize provided by the researcher, and that the student submitting the best project would also receive a prize. Upon completion of the study, two portable CD players were sent to two students, one to a student who was randomly selected from all students who completed their projects and one to the author of the best project judged by the researchers.

Treatment conditions. Each teacher taught two classes. All experimental groups completed the *Learning Explorations* program with variations described below.

Students in the self-evaluation group conducted their experiments about learning, produced written reports, and formally self-evaluated their work at two times during the instruction. Students in this group used the Project Rating Scale to self-evaluate their initial method section before conducting the experiment, and their initial final report before revising and finalizing it. They checked the most appropriate description for their project for each item on the rating scale and wrote their ideas for improvement. Students could then revise their initial method section and initial report to incorporate their ideas for improvement into their final reports. These students received no teacher evaluation on their reports.

Students in the teacher-evaluation group conducted their experiments about learning and produced written reports, and they received written teacher evaluation on the Project Rating Scale at the same two times during the instruction as did the students in the self-evaluation condition. Students could then revise their work. The teachers were provided with sample feedback statements for use at their discretion, such as "Specify how the dependent variable will be measured," "State the hypothesis in 'If . . . then . . .' format," or "Describe how the treatment groups will be different." The teachers received training on the use of the Project Rating Scale during the 4-hr teacher-training workshop described further in this paper. In order to minimize potential cognitive overload for students, the teachers were told to comment primarily on the lowest-rated items, and not to exceed five comments for any student.

Students in the self-plus-teacher evaluation group conducted their experiments about learning and produced written reports, and formally self-evaluated their work two times during the instruction in the same manner as the self-evaluation group. After the students had completed each of their two self-evaluations, they submitted to the teacher their initial method section or their report and the Project Rating Scale with their self-evaluation. The teacher then wrote an evaluation on the rating scale along with the student self-evaluation by checking the most appropriate description for the student project on each criterion, and by writing suggestions for improvement in the same manner as in the teacher-evaluation condition. The teachers received the same sample feedback statements for use at their discretion as the teachers in the teacher-evaluation condition. The students could then revise their work based on their own self-evaluation and on the teacher evaluation.

Students in the no-evaluation, or control, group conducted their experiments about learning and produced written reports. Similarly to the other three treatment groups, they were provided with the Project Rating Scale before they began work on their experiments. They were given the same amount of time as the other groups to conduct the experiments and write their reports, and could revise their work as much as they found necessary. They received no formal feedback from the teacher, and they were not asked to formally evaluate their own work. Thus, the main difference between the no-evaluation and the self-evaluation group rests in the fact that the no-evaluation group did not engage in formal self-evaluation procedure by applying the Project Rating Scale to their initial work during class time.

Each teacher taught the program during regular classes over approximately six weeks, two 40-min class periods per week. One teacher in each treatment taught the program as part of a high school psychology course; the other taught it as part of another course.

Students took the posttest during the next-to-last class period of the treatment. In the final class period, they submitted their experiment reports and completed an attitude survey about the program. A diagram summarizing the experimental design is shown in Table 3.

Time in program. The teachers reported spending an average of 12.5 forty-min class periods on the instructional program, with a range of 12 to 14 class periods. Teachers spent an average of 13 class periods on the program in the no-evaluation group, 12.5 periods in the teacher-evaluation and the self-plus-teacher evaluation groups, and 12 periods in the self-evaluation group.

Classroom observations and focus groups. During implementation of the instructional program, two trained and experienced observers con-

Table 3 ☐ Summary of experimental design.

Assignment of Teachers to Treatment Groups
(Two teachers per treatment group, two classes per each teacher)

↓

Treatments

No Evaluation	*Self-Evaluation*	*Teacher-Evaluation*	*Self + Teacher Evaluation*
• Students receive instruction on experimental design components.	• Students receive instruction on experimental design components.	• Students receive instruction on experimental design components.	• Students receive instruction on experimental design components.
• Students design experiments.	• Students design experiments.	• Students design experiments.	• Students design experiments.
• Students receive training on use of project rating scale.	• Students receive training on use of project rating scale.	• Students receive training on use of project rating scale.	• Students receive training on use of project rating scale.
	• Students self-evaluate method section.	• Teacher evaluates method section.	• Students self-evaluate and teacher evaluates method section
• Students receive instruction on writing reports.	• Students receive instruction on writing reports.	• Students receive instruction on writing reports.	• Students receive instruction on writing reports.
• Students conduct experiments and write initial reports.	• Students conduct experiments and write initial reports.	• Students conduct experiments and write initial reports.	• Students conduct experiments and write initial reports.
	•Students self-evaluate initial reports.	• Teacher evaluates initial reports	Students self-evaluate and teacher evaluates initial reports.
• Students produce final reports.	• Students produce final reports.	• Students produce final reports.	• Students produce final reports.

↓

Posttest

↓

Attitude Survey (Student)
Students Submit Experiment Reports

↓

Attitude Survey (Teacher)

ducted a classroom observation of each of the eight teachers' classrooms to ensure that the procedures in the teacher guide were closely followed and that the treatments were properly implemented. Both observers were former classroom teachers and school administrators with experience in delivery of inservice training to teachers. The observers used an eight-item Classroom Observation Checklist to focus and record their observations. Observer ratings of program delivery during the 10 classroom observation visits were generally favorable with a mean rating of 2.19 on a 4-point scale with a highest possible score of 3. The observers reported that the teachers were well prepared for the lessons, that the students were given accurate information about how to plan and conduct experiments, and that the experimental procedures were carried out properly.

The two observers also conducted focus group meetings of approximately 30 min each, at the end of the program, with one class of students in each treatment condition. The students were asked about their likes and dislikes regarding the instructional program and their attitudes toward the two evaluation strategies.

Design and Data Analysis

This was a quasi-experimental posttest-only control-group design. No pretest was used because the instructional program focused on content that was quite unfamiliar to the students. In addition, the experimenters believed that pretesting students on such unfamiliar content might be frustrating to them and, consequently, result in negative attitudes toward the instructional program. The data analysis for student performance was carried out as a one-way multivariate analysis of variance (MANOVA) with four groups (no-evaluation, self-evaluation, teacher evaluation, and self-plus-teacher evaluation) and two dependent variables, the student project scores and posttest scores.

Attitude data on the six items common to all four groups were analyzed using a MANOVA, followed by univariate ANOVAs where appropriate. Mean scores for the additional survey items for the last three treatment conditions were analyzed separately. The differences between student attitudes toward self-evaluation and teacher evaluation for students in the self-evaluation, teacher-evaluation, and self-plus-teacher evaluation groups were explored using independent-samples t tests and paired-samples t tests.

RESULTS

The results for achievement, time in program, student attitudes, teacher attitudes, analysis of student initial reports, and classroom observations are reported in this section.

Achievement

The mean scores and standard deviations for both the project reports and the posttest are shown in Table 4. The table reveals that the mean scores for the project reports were 17.55 (59%) for the no-evaluation group, 18.35 (61%) for the self-evaluation group, 20.99 (70%) for the teacher-evaluation group, and 21.44 (72%) for the self-plus-teacher evaluation group. The mean posttest scores were 21.56 (72%) for the no-evaluation group, 22.67 (76%) for the self-evaluation group, 23.11 (77%) for the teacher-evaluation group, and 22.48 (75%) for the self-plus-teacher evaluation group.

The overall mean scores were 19.54 (65%) for the project reports and 22.45 (75%) for the posttest. The Pearson correlation coefficient between the overall project scores and the posttest scores was .41.

The one-way MANOVA conducted on the project report and posttest scores yielded a significant overall difference, Wilks's $\Lambda = .891$, $F(6, 672) = 6.63$, $p < .001$, $\eta^2 = .056$. Follow-up univariate tests revealed significant differences between treatment groups on the project scores, $F(3, 337) = 11.85$, $MSE = 26.33$, $p < .001$, $\eta^2 = .095$, but not on the posttest scores, $F(3, 337) = 1.92$, $MSE = 17.65$, $p = .127$, $\eta^2 = .017$.

Post hoc tests of the project score differences were conducted between treatment groups. Using the Bonferroni adjustment, each of the six pairwise comparisons was tested at .008 level (.05/6 tests). These comparisons revealed that the teacher-evaluation group had significantly

Table 4 ☐ Mean project report and posttest scores by treatment.

	Treatment				
Measure	*No evaluation*	*Self-evaluation*	*Teacher evaluation*	*Self + teacher evaluation*	*Total*
Project Reports					
M	17.55	18.35	20.99	21.44	19.54
SD	(6.21)	(5.11)	(4.75)	(4.29)	(5.37)
Posttest					
M	21.56	22.67	23.11	22.48	22.45
SD	(3.75)	(4.58)	(4.28)	(4.10)	(4.22)

Note. The maximum possible score on both the project reports and the posttest was 30.

higher scores on their project reports than the no-evaluation group ($p < .001$) and the self-evaluation group ($p = .006$), and that the self-plus-teacher evaluation group had significantly higher project report scores than the no-evaluation group ($p < .001$) and the self-evaluation group ($p < .001$). The differences in the project report scores between the no-evaluation and the self-evaluation groups and between the teacher-evaluation and the self-plus-teacher evaluation groups were not statistically significant.

Student Attitudes

Basic attitude survey. The mean attitude scores by treatment for the student responses to the six statements on the 4-point Likert-type attitude survey administered after completion of the instructional program are shown in Table 5. The overall mean score across the six Student Attitude Survey items was 1.92, a moderately favorable rating indicating general agreement with positive statements about the instructional program. The two highest-rated items on the survey were "I am satisfied with my experiment and the experiment report" ($M = 2.12$) and "I can now write a research report" ($M = 1.97$). The two lowest-rated items were "I liked planning and conducting experiments about learning" ($M = 1.76$) and "I liked the program" ($M = 1.78$).

The data in Table 5 were analyzed using a 4 (Treatment) × 6 (Survey Items) MANOVA to test for significant differences. The overall means were significantly different across the four treatment groups, Wilks's $\Lambda = .789$, $F(18, 877) = 4.26$, $p < .001$. The strength of the relationship between the treatments and student attitude scores, as measured by η^2, was moderate, with the treatments accounting for 8% of the variance of student attitude scores.

Follow-up ANOVAs conducted on each of the six items revealed significant attitude differences at or below the .001 level between the treatment groups on five of the items. Post hoc tests for the pairwise comparisons on these five items revealed 14 significant differences between groups. The two groups containing self-evaluation, the self-evaluation group and the self-plus-teacher evaluation group, had the most positive attitudes. Students in both these groups reported learning from the program more, being more satisfied with their experiment, and being more confident in their ability to independently conduct experiments and to write research reports than did the no-evaluation group. Students in the self-evaluation group also reported liking the program more than did students in the no-evaluation group.

In addition to having more positive attitudes than the no-evaluation group, the two groups with the self-evaluation condition had significantly more positive attitudes than the teacher-evaluation group on several items. Students in both the self-evaluation and the teacher-plus-self evaluation groups reported learning from the program more than students in the teacher-evaluation group. Students in the self-evaluation group also indicated that they liked the program more, and that they felt more confident

Table 5 ☐ Mean student attitude scores by treatment.

Item	*No evaluation*	*Self-evaluation*	*Teacher evaluation*	*Treatment Self + teacher evaluation*	*Means*	*F*	*p*
1. I liked the program.	1.64[2]	1.93[1]	1.62[2]	1.89	1.78	6.03	<.001
2. I learned a lot . from the program	1.68[2]	2.16[1]	1.78[2]	2.12[1]	1.95	11.90	<.001
3. I liked planning and conducting experiments about learning.	1.62	1.89	1.72	1.77	1.76	1.93	*ns*
4. I am satisfied with my experiment and the experiment report.	1.82[2]	2.21[1]	2.16[1]	2.28[1]	2.12	8.87	<.001
5. I can now independently design and conduct experiments about learning.	1.74[2]	2.13[1]	1.86[2*]	2.06[1]	1.95	9.04	<.001
6. I can now write a research report.	1.71[2]	2.05[1]	1.94	2.16[1]	1.97	10.82	<.001
Overall means	1.70	2.06	1.85	2.05	1.92	4.26	<.001

Note. All items were measured on a four-point scale: 0 (*strongly disagree*), 1 (*disagree*), 2 (*agree*), 3 (*strongly agree*).
1. & 2. Each score with a superscript of one for an item is significantly more positive than those marked with a two.
* The one exception is on item five, where the self-evaluation group but not the self-plus-teacher evaluation group was significantly more positive than the teacher-evaluation group.

Table 6 ☐ Ratings of additional student survey items regarding self-evaluation and teacher evaluation.

Item	*No evaluation*	*Self-evaluation*	*Treatment Teacher evaluation*	*Self + teacher evaluation*	*Means*
7. I liked evaluating my written report.[a]	–	1.59	–	1.61	1.60
8. My evaluation of my report helped me improve it.	–	1.91	–	2.18	2.04
9. I liked receiving teacher feedback on my written report.	–	–	1.99	2.35	2.19
10. Teacher feedback on my report helped me improve it.	–	–	2.40	2.68	2.55
Overall means	–	1.75	2.20	2.21	2.10

Note. All items were measured on a four-point scale: 0 (*strongly disagree*), 1 (*disagree*), 2 (*agree*), 3 (*strongly agree*).a Item numbers correspond to their numbers on the Student Attitude Survey.

in their ability to independently conduct experiments than did students in the teacher-evaluation group.

Additional attitude survey items. The mean scores by treatment for additional student survey items regarding self-evaluation and teacher evaluation are shown in Table 6. The mean score on these items regarding self-evaluation was 1.75 for the self-evaluation group, and for these two items regarding teacher evaluation was 2.20 for the teacher-evaluation group. The mean score for the four items regarding both self- and teacher evaluation was 2.21 for the self-plus-teacher evaluation group.

Independent-sample *t* tests revealed that the self-plus-teacher evaluation group had significantly more positive attitudes toward self-evaluation than the self-evaluation group at the .01 level on the item "My evaluation of my report helped me improve it." The self-plus-teacher evaluation group also had significantly more positive attitudes toward teacher evaluation than the teacher-evaluation group at .01 level on the items "I liked receiving teacher feedback on my written report" and "Teacher feedback on my report helped me improve it."

We used *t* tests to compare the attitudes of students in the self-plus-teacher evaluation group toward the two evaluation strategies. These tests revealed that students rated each of the teacher evaluation items ("I liked receiving teacher feedback on my written report" and "Teacher feedback on my report helped me improve it") significantly higher at the .001 level than the respective self-evaluation items ("I liked evaluating my written report" and "My evaluation of my report helped me improve it").

Focus groups. Students in the focus groups were questioned in some detail about their attitudes toward teacher evaluation and self-evaluation. Students in the teacher-evaluation group found teacher evaluation helpful in improving their work. This idea is captured in the following student comment: "Yes. Teacher evaluation indicated exactly what I did not have. It showed exactly what I needed to add. And then I could write it." Students in the teacher-evaluation group also noted that, in instruction on other topics, they rarely had an opportunity to receive teacher feedback on their work before getting a grade. They reported that, during their regular classes, most often the teacher explained common mistakes to the class after student work had already been graded.

Student opinions about self-evaluation differed. Most students thought that it was difficult for them to evaluate their own work. Students felt that they were not as objective and strict in evaluating their own work as a teacher might be, as indicated by the following student comment: "I think that it is very difficult to evaluate oneself. When writing your work you already think that it is the best. And you write it in a way that you find the best. I don't know. It is very hard." Several students also thought that they were not able to see their own mistakes as well as someone else would and that they lacked prior experience in evaluating their work. A few students thought that they would be able to evaluate their work, especially if given clear evaluation criteria.

Teacher Attitudes

The overall mean ratings by treatment to the eight Likert-type items on the Teacher Attitude Survey were 1.88 for the no-evaluation group, 2.19 for the self-evaluation group, 2.25 for the teacher-evaluation group, and 2.06 for the self-plus-teacher evaluation group. The two highest-rated items were "The program was appropriate for my students" (M = 2.50) and "I would use the program again with my students" (M = 2.43). The two lowest-rated items were "I felt like I did a good job of teaching the program" (M = 1.63) and "The program was easy to teach" (M = 1.88). No statistical test was conducted of the differences between mean attitude scores by treatment because there were only two teachers for each treatment.

Like students, teachers also rated teacher evaluation items higher than the self-evaluation items. All four teachers in the two groups with the teacher-evaluation component agreed or strongly agreed that providing written feedback on student draft reports was too time consuming.

Student and Teacher Ratings of Initial Research Projects

Student and teacher ratings of initial method sections and initial research reports were also examined. Students in both the self-evaluation and self-plus-teacher evaluation groups tended to considerably overrate the quality of their initial reports compared to the experimenter ratings. The mean student rating of their *initial* reports was 26.69 for the self-evaluation group and 27.36 for the self-plus-teacher evaluation group, out of a maximum of 30 points, while the mean experimenter ratings of the *final* version of the same reports were 18.85 and 21.36 respectively. Students in both groups with the self-evaluation component rated their initial reports higher in more than 90% of the cases than the experimenter rated their final reports.

Students in both groups with the self-evaluation component assigned lower-than-maximum ratings of 2 on an average of only 2 of the 15 criteria on the initial report. Student reports for the self-evaluation groups improved from the initial to the final reports in these lower-rated criteria in approximately 45% of the cases. For the most part, students rated their projects without providing comments regarding specific ideas for improvements. In several cases, students struggled to identify what improvements were necessary, as indicated by comments such as "It is difficult to evaluate my own work, because to me it seems that everything is always correct" and "I think that something is missing, but do not know what."

Teacher ratings of the initial student reports were more consistent with the experimenter ratings of the final reports than were the student ratings. The mean teacher rating of student initial reports was 20.45 for the teacher-evaluation group and 23.69 for the self-plus-teacher evaluation group, while the mean experimenter ratings of the final version of the same reports were 21.11 and 21.54 respectively.

Teachers in both groups with the teacher-evaluation component assigned lower-than-maximum possible ratings on the initial report on an average of 5 of the 15 criteria, whereas students had assigned lower-than-maximum ratings on an average of only 2 criteria. Student final reports for both groups with the teacher-evaluation component improved from the initial to the final reports in these lower-rated areas in approximately 65% of the cases. When there was a discrepancy between student and teacher ratings in the self-plus-teacher evaluation treatment, students appeared to rely primarily on the teacher evaluation. However, when students did not improve their work as a result of teacher comments, it appeared that they had not learned the relevant concepts well enough to know how to make improvements from the teacher comments.

DISCUSSION

In this study, we examined the effects of student self-evaluation, teacher evaluation and the combination of self-evaluation and teacher evaluation on student performance and attitudes. Students in the teacher-evaluation and the self-plus-teacher evaluation conditions received significantly higher experimenter ratings on their research projects than students in the no-evaluation and the self-evaluation conditions. There were no significant differences between the treatment groups on the posttest. Overall, students in both groups with the self-evaluation component reported more positive attitudes toward the program on the attitude survey and had greater confidence in their ability to conduct experiments in the future than students in the teacher-evaluation and the no-evaluation groups. However, when specifically asked about their attitudes toward the two evaluation strategies on the attitude survey and during focus groups, both students and teachers indicated that they liked teacher evaluation better and thought that it improved student research projects to a greater extent than student self-evaluation.

Teacher Evaluation and Project Performance

The better student performance on research reports under teacher evaluation appears to be due to better evaluation and feedback provided by the teachers. Analysis of the differences between student and teacher ratings of the initial research projects showed that the teachers

applied the Project Rating Scale to the student projects more reliably and interpreted the scale using higher standards than did the students. Teachers assigned more lower-than-maximum ratings than did students on more than twice as many criteria on the Project Rating Scale. Teachers also provided more specific suggestions for improvement of the reports than students did in their self-evaluations. In addition, students incorporated teacher suggestions more frequently than their own evaluations in their final projects. Students may have found teacher suggestions easier to incorporate because these provided more concrete suggestions on what they should do.

Student Self-Evaluation and Project Performance

The study yielded strong evidence that students lacked ability and confidence to evaluate their own work. More than 90% of students rated their initial projects considerably higher than the experimenter rated their final projects, thus reducing the possibility of improving their own work. There are several plausible explanations for the lack of precision in student ratings of their own projects. First, as indicated by several students in the attitude survey, students were not familiar with self-evaluation because it is rarely used in Latvian schools. Systematic or precise self-evaluation may not be common in other countries either. Students in most classrooms come to rely on the teacher as the sole assessor of their performance. As a result, student opportunities for development of self-assessment skills are hampered. Yet, regular and frequent opportunities for engaging in self-evaluation is thought to increase the accuracy of student self-evaluations (Fontana & Fernandez, 1994; Ross et al., 1999; Sadler, 1989).

Secondly, student lack of confidence in self-evaluation and their dislike of it may have led them to invest little effort in their self-evaluations. Several students in attitude surveys and focus groups indicated that they did not like evaluating their own work, that they found it difficult, and that they could not be objective. Third, some students may have purposely assigned higher ratings for their projects to boost their confidence and good feelings about their work and to look better in front of the teacher. Several focus group participants noted that receiving a good grade, or more points in this case, gave students a better perception of their own work. This explanation reflects a self-presentation bias noted by other researchers (Shen, Sullivan, Igoe, & Shen, 1996).

It is likely that many students did not learn the content of the instructional program well enough to fully understand and internalize the standards on which they were to assess their reports and to identify the most appropriate strategies for improving them. A number of students during focus groups and on their self-evaluations either indicated that they were unable to detect what they needed to improve or reported that something did not seem quite right in their report but they could not determine what. In addition, even though students in both teacher evaluation conditions were provided with improvement suggestions from the teacher, student reports improved from the initial to the final version in only about 65% of the cases when such suggestions were provided.

Self-Evaluation and Student Attitudes

Students in both groups with the self-evaluation component had more positive atitudes toward the program on most of the survey items than students in the other two groups. Formal self-evaluation may have caused students to think that they knew the criteria better for designing and reporting a research project. They may have felt more in control of their learning than did students in the other two groups who did not evaluate their own work. This interpretation is consistent with findings of other researchers that students who self-monitor and self-evaluate their progress have more positive attitudes toward learning and higher self-efficacy perceptions than those who do not (Schunk, 1989; Zimmerman & Kitsantas, 1999). In addition, the experience of applying the Project Rating Scale to their own work may have caused students in the self-plus-teacher evaluation group to be more prepared to accept teacher suggestions for improvement and to view teacher evaluation as being based on more objective criteria.

Posttest Performance

The lack of significant differences on the posttest was most likely due to the fact that teacher evaluation and student self-evaluation were applied to the design and write-up of the research projects, but not directly to the posttest content. Producing the research projects was a more difficult and advanced task than performing well on the posttest, so the former task was the focus of the self-evaluation and teacher-evaluation strategies. The more challenging nature of the research projects is also demonstrated by the fact that the mean scores on the student projects were lower than on the posttest, with overall means of 19.54 out of 30 possible (65%) for the research projects and 22.45 out of 30 possible (75%) for the posttest.

Future Research

Although the study took place in Latvia, its results are relevant to the use of self- and teacher evaluation in the United States and other countries. The study confirmed the important role of formative teacher feedback in improving student performance by demonstrating the positive effects of teacher evaluation on performance of a complex learning task. In addition, the study also confirmed earlier findings that student engagement in self-evaluation may lead to increased self-efficacy perceptions resulting in greater confidence to pursue similar tasks in the future. However, the study also raised several issues that should be addressed.

Conducting the study in Latvia, where teachers and students are quite unfamiliar with formative teacher evaluation and student self-evaluation, highlighted the importance of the larger teaching and learning context in which these formative classroom assessment strategies are embedded. Several authors have noted that use of formative assessment strategies, and student self-assessment in particular, requires a change in roles for both teachers and students (Black, 1993, Gipps, 1994; Shepard, 2000; Stiggins, 2001). Further research could help us determine how to better implement effective formative assessment strategies in the classroom.

The effects of frequent practice opportunities on student ability to evaluate their own work also deserves further investigation. Researchers could provide teachers with instruction on designing self-evaluation opportunities around existing classroom tasks in different subject areas. Student abilities to self-evaluate, as well as their attitudes toward self-evaluation, could be tracked over time or compared with those of students who do not receive self-evaluation practice. In such longer-term studies, one could also examine whether student self-evaluation abilities would transfer from one task to the next within a subject domain and across domains.

Further research on the nature of such self-evaluation practice, especially concerning complex performance-based tasks such as the one used in this study, is also warranted. Sadler (1989) suggested that, in order for the students to hold a concept of quality roughly similar to that of the teacher, students should be presented with standards and multiple exemplars and engage in direct evaluative experiences. He also recommended peer evaluation as a way to develop student self-assessment skills and to help students acquire strategies for improving their work. Research is necessary to determine the role of each of these components in the development of student self-evaluation ability, the optimum extent of such training, and its effects on student performance and attitudes.

Finally, teacher and student attitudes toward teacher evaluation and student self-evaluation should be analyzed in a greater depth.This study indicates that, even though neither students nor teachers favored self-evaluation, it did raise the confidence of students in their ability to conduct experiments and write research reports. Teacher evaluation, on the other hand, yielded less-positive student attitudes. Further exploration into the reasons for the student and teacher attitudes toward the two evaluation strategies could help to determine the most appropriate procedures for using them in the classroom. □

Zane Olina [olina@coe.fsu.edu] is Assistant Professor in Instructional Systems Program at Florida State University. Howard Sullivan is Professor in Educational Technology Program at Arizona State University.

This research is based on Dr. Olina's dissertation at Arizona State University, and was supported by

grants from the Arizona State University Graduate Research Support Program and the Assessment Training Institute Foundation in Portland, Oregon. We gratefully acknowledge their support.

REFERENCES

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review, 84,* 191–215.

Black, P. (1993). Formative and summative assessment by teachers. *Studies in Science Education, 21,* 49–97.

Black, P., & William, D. (1998a). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7–75.

Black, P., & William, D. (1998b). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan, 80*(2), 139–144.

Bransford, J. D., Brown, A. L., & Cocking, R. R. (Eds.). (1999). *How people learn: Brain, mind, experience, and school.* Washington D.C.: National Academy Press.

Butler, R. (1988). Enhancing and undermining intrinsic motivation: The effects of task-involving and ego-involving evaluation on interest and performance. *British Journal of Educational Psychology, 58,* 1–14.

Cardelle-Elawar, M., & Corno L. (1985). A factorial experiment in teachers' written feedback on student homework: Changing teacher behavior a little rather than a lot. *Journal of Educational Psychology, 77,* 162–173.

Ertmer, P. A., & Newby, T. J. (1996). The expert learner: Strategic, self-regulated, and reflective. *Instructional Science, 24,* 1–24.

Flavell, J. H. (1976). Metacognitive aspects of problem solving. In L. Resnick (Ed.), *The nature of intelligence* (pp. 231–236). Hillsdale, NJ: Lawrence Erlbaum Associates.

Fontana, D., & Fernandez, M. (1994). Improvements in mathematics performance as a consequence of self-assessment in Portuguese primary school pupils. *British Journal of Educational Psychology, 64,* 407–417.

Gipps, C.V. (1994). *Beyond testing: Towards a theory of educational assessment.* London: The Falmer Press.

Hughes, B., Sullivan, H. J., & Mosley, M. L. (1985). External evaluation, task difficulty, and continuing motivation. *Journal of Educational Research, 78,* 210–215.

Klenowski, V. (1995). Student self-evaluation processes in student-centred teaching and learning contexts of Australia and England. *Assessment in Education: Principles, Theory & Practice, 2*(3), 145–163.

Maqsud, M., & Pillai, C. M. (1991). Effect of self-scoring on subsequent performances in academic achievement tests. *Educational Research, 33,* 151–154.

Olina, Z., & Sullivan, H. J. (2002). Effects of classroom evaluation strategies on student achievement and attitudes. *Educational Technology Research and Development, 50*(3), 61–75.

Page, E. B. (1958). Teacher comments and student performance: A seventy-four classroom experiment in school motivation. *Journal of Educational Psychology, 49,* 173–181.

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

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

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18,* 119–144.

Salili, F., Maehr, M. L., Sorensen, R. L., & Fyans, L. J., Jr. (1976). A further consideration of the effects of evaluation on motivation. *American Educational Research Journal, 13,* 85–102.

Schunk, D. H. (1989). Social cognitive theory and self-regulated learning. In B. J. Zimmerman & D. H. Schunk (Eds.), *Self-regulated learning and academic achievement: Theory, research and practice* (pp. 83–110). New York: Springer-Verlag.

Shen, S., Sullivan, H., Igoe, A., & Shen, X. (1996). Self-presentation bias and continuing motivation among Chinese students: A cross-cultural phenomenon. *Journal of Educational Research, 90* (1), 52–56.

Shepard, L. A. (2000, October). The role of assessment in a learning culture. *ER Online* [Online], *28.* Available http://www.aera.net/meeting/am2000/wrap/praddr10.htm

Stewart, L. G., & White, M.A. (1976). Teacher comments, letter grades, and student performance: What do we really know? *Journal of Educational Psychology, 68,* 488–500.

Stiggins, R. J. (2001). *Student-involved classroom assessment.* Upper Saddle River, NJ: Merrill Prentice Hall.

Thomas, J. W., Bol, L., Warkentin, R. W., Wilson, M., Strage, A., & Rohwer, W. D., Jr. (1993). Interrelationships among students' study activities, self-concept of academic ability, and achievement as a function of characteristics of high-school biology courses. *Applied Cognitive Psychology, 7,* 499–532.

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

Winne, P. H., & Hadwin, A. F. (1998). Studying as self-regulated learning. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Metacognition in educational theory and practice* (pp. 277–304). Makwah, NJ: Lawrence Erlbaum Associates.

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

Zimmerman, B. J. (1989). A social cognitive view of self-regulated academic learning. *Journal of Educational Psychology, 81,* 329–339.

Zimmerman, B. J., & Kitsantas, A. (1999). Acquiring writing revision skill: Shifting from process to outcome self-regulatory goals. *Journal of Educational Psychology, 91,* 241–250.