# Effects of Classroom Evaluation Strategies on Student Achievement and Attitudes

Zane Olina
Howard J. Sullivan

*This study investigated the effects of teacher evaluation and the combination of teacher evaluation and student self-evaluation on student performance and attitudes. Participants in the study were 189 Latvian high school students and their six teachers.*

*The six teachers were assigned to one of three treatment conditions: (a) no evaluation, (b) teacher evaluation, and (c) self-evaluation plus teacher evaluation. All groups completed a 12-lesson instructional program on how to conduct experiments and produce research reports. Students in the teacher-evaluation group received teacher evaluation on their initial research reports. Students in the self-plus-teacher evaluation group self-evaluated their reports and received teacher evaluation on them. The no-evaluation group received no formal evaluation instructions.*

*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 research experiments.*

Classroom evaluation can have a powerful impact on student performance and motivation (Crooks, 1988; Natriello, 1987). In his review, Crooks cited evidence that evaluation can provide students with knowledge of results and corrective feedback, help students monitor their own progress, and influence students' continuing motivation and their perceptions of their self-efficacy as learners. From a review of 250 articles on classroom assessment, Black and William (1998) reported positive effects of formative evaluation on performance of students of all ages and ability levels. They used the term *formative evaluation* to refer to the evaluation of instruction for the purpose of improving student performance, rather than to evaluation of an instructional program for the purpose of improving it during its developmental or formative stage (Scriven, 1967).

Teacher evaluation of student work is the most common form of classroom evaluation cited as having positive effects on student performance and attitudes. Cardelle-Elawar and Corno (1985) found that elementary school student performance and attitudes toward mathematics improved when their 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.

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. Story and Sullivan (1986) found that teacher comments had no significant main effects on the continuing motivation of fifth- and sixth-grade students, but the combination of comments and an easier task were effective in motivating girls to return to the same task.

Researchers have cited several characteristics of effective teacher evaluation. Students should be provided with explicit evaluation criteria and models of good work before they begin working on a learning task (Sadler, 1989; Wiggins, 1998). Teacher evaluation of performance is most effective when it is specific, directly related to the task, and provides the opportunity for students to correct their performance (Black & William, 1998; Crooks, 1988). Finally, teacher evaluation should be used initially for formative purposes rather than for assigning a grade (Hughes, Sullivan, & Mosley, 1985; Sadler, 1989).

Student self-evaluation is a second form of classroom evaluation that may enhance student learning. There has been considerable emphasis in the last decade on active student evaluation of their own work (Bransford, Brown, & Cocking, 1999; Gipps, 1994; Wolf, Bixby, Glenn, & Gardner, 1991). Several authors have argued that student self-evaluation has the potential to improve performance and motivation (Gipps, 1994; Shepard, 2000; Wiggins, 1998). Yet evaluation by students of their own work is not a common practice in the classroom, and the topic is frequently overlooked in general literature on classroom assessment (Black & William, 1998).

Overall, studies show positive effects of self-evaluation on student performance and motivation across subject areas and age groups. Maqsud and Pillai (1991) reported 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 their teacher. In a yearlong intervention in Portugal, Fontana and Fernandez (1994) found that primary school students trained in self-evaluation performed better in mathematics than students who did not receive self-evaluation training. Ross, Rolheiser and Hogaboam-Gray (1999) reported similar results with fourth- to sixth-grade students who were trained in self-evaluation of narrative writing compared to their counterparts who were not trained in self-evaluation.

Studies involving comparisons of teacher evaluation and student self-evaluation have shown positive effects for self-evaluation on student continuing motivation. Salili, Maehr, Sorensen, and Fyans (1976) had fifth-grade Iranian students solve word anagram problems under three different evaluation conditions. Students in the teacher-evaluation condition were told that their work would be evaluated by the teacher, students in the self-evaluation condition were told that no one would know about their results, while students in the peer-comparison condition were told the results would enable them to see how well they did on the task in comparison with their classmates. Salili et al. found no performance differences among the three groups. Yet students in the self-evaluation and peer-comparison conditions showed greater continuing motivation than students in the teacher-evaluation condition in the form of desire to return to the initial task. Hughes et al. (1985) reported that students returned to a difficult task more often after self-evaluation and to an easy task more often after teacher evaluation.

Research findings on the use of teacher evaluation and student self-evaluation to improve student learning prompted the first author of this article to investigate the potential for these instructional strategies in her home country, Latvia. 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. The National Compulsory Education Standard of the Latvian Ministry of Education and Science Center for Curriculum Development and Examination (LMES CCDE, 1998), as well as the Subject Content Standards issued by the LMES CCDE, emphasize that both teacher evaluation and student self-evaluation are to be used as an integral part of the teaching and learning process.

Most Latvian teachers are not prepared for the new expectations, and very little profes-

sional development is provided for improving teacher evaluation skills. Under the Russian regime, all schools used the same textbooks and a prescribed curriculum. Because of the education reform efforts in the past decade, only Subject Content Standards have been mandatory for schools, and teachers have been free to design their own curricula and assessment approaches. The authoritarian classroom management style by teachers still dominates in most classrooms, and there is no tradition of engaging students in evaluation of their own work. Because of large workloads with almost half of the teachers teaching more than 32 direct contact hours a week (Soros Foundation-Latvia, 2001), many teachers also lack the time and experience to provide students with constructive formative feedback about their performance.

The present study investigated the effects of teacher evaluation and student self-evaluation on student posttest scores, the quality of student research reports, and student attitudes. Three levels of evaluation were employed: (a) a no-evaluation or control condition; (b) a teacher-evaluation condition, and (c) a self-plus-teacher evaluation condition. The study was conducted in the students' regular classes, and their regular teachers delivered the instructional treatments.

The teacher-evaluation component in this study was designed based on several teacher-evaluation strategies reported as effective by authors of evaluation research (Black & William, 1998; Hughes et al., 1985; Sadler, 1989; Wiggins, 1998). Students were provided 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 teacher received. The students then applied the rating 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).

The following research questions were investigated:

1. Does teacher evaluation have a positive effect on student performance?
2. Does the combination of teacher evaluation and student self-evaluation have a different effect on student performance than teacher evaluation alone?
3. Does the combination of teacher evaluation and student self-evaluation have a different effect on student attitudes than teacher evaluation alone?

The research questions dealing with student performance were investigated using two different criterion measures, (a) scores on a posttest covering the instructional content and (b) researcher ratings of the student research reports. The use of the research report measure is consistent with the recommendation that students can best develop self-evaluation skills when they are given authentic performance-based learning tasks (Stiggins, 2001; Wiggins, 1998; Wolf et al., 1991). The instructional program and all assessment instruments used in this study were field tested earlier by the researcher with a similar age group of students to the target population.

## METHOD

### Participants

Participants in the study were 189 Latvian high school students from 12 classes taught by six teachers. The average class size was 16 students. All teachers involved in the study had completed at least four years of college and obtained either a bachelor's or master's degree in education. The teachers represented a variety of subject areas, such as language arts, mathematics,

science, and social studies. The classes were drawn from five schools in different regions of Latvia, representing both rural and urban areas and varied socio-economic backgrounds. One school was located in a city of 100,000 people, three schools were in towns of approximately 10,000, and one was in a small town of approximately 4,000.

## Materials

A 12-lesson instructional program entitled *Learning Explorations* was developed in print form in Latvian for use in the study. The purpose of the program was to introduce high school students to the basic concepts of scientific research and to develop their skills in designing experiments about learning. 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. *Learning Explorations* was designed to teach the following six objectives, which are listed here in an abbreviated form. Students were (a) to identify the major experimental design concepts (hypothesis, dependent variable, control group, independent variable, treatment conditions, constants), (b) to indicate these concepts in an experiment scenario, (c) to detect common experimental design flaws in an experiment scenario, and (d) to write a set of experimental procedures and (e) summarize the results for an experiment scenario. The sixth objective required students (f) to independently conduct an experiment about learning and to produce a simple written report of the results.

The first part of the program introduced students to the basic experimental design concepts and common experimental design flaws. Students were then provided with a problem scenario of a fictional high school in Latvia that had recently received funding for purchase of new textbooks. They were asked to assist the school's textbook committee in developing criteria for selecting the best textbooks. Students were provided with two example experiments, including complete experimental procedures and instructions for participants, and were told that they could conduct one of the two given experiments or design their own. The great majority of students chose to conduct the two experiments provided in the program. They had to have at least five participants in each treatment group. Students had to report their experiments using a two-page experiment report form that contained spaces for introduction, method, results, conclusions and references sections that were broken down in several subsections. The last part of the instructional program provided students with detailed instruction on how to complete the experiment report form.

The *Learning Explorations* program was organized into six sections, each of which required about two 40-min class periods. The program materials consisted of a student book and a teacher guide. The student book contained all the information presented during instruction, and examples of experiments, practice exercises, and worksheets. The teacher guide included step-by-step lesson plans on how to use the student book, descriptions of instructional activities, a posttest, a rating scale for student projects, and transparencies and handout masters. All materials used in the study were in the Latvian language.

## 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.* During the program, students received a two-page report form on which to describe the results of their experiments. An independent rater trained by the researcher evaluated all student projects on the project rating scale, a descriptive rating scale for evaluating written research reports developed for the study. The rater was a former classroom teacher with considerable experience rating student projects. The researcher rated one sixth of all student projects to determine inter-rater reliability. The raters were blind to the experimental condition and the school of origin for each project. The Pearson correlation coefficient

for the inter-rater reliability between the ratings of the final student projects by the independent rater and the researcher was .90.

The project rating scale consisted of 15 items rated on a three-point scale from 0 to 2. The 15 criteria included in the project rating scale are listed in Figure 1. A sample item from the project rating scale is provided below. Two points were assigned for above average performance and 0 points for below average performance, as shown in the example below:

*Describes experimental procedures step-by-step.*

❒ describes experimental procedures step-bystep, including all materials and conditions (2 points)

❒ some steps and conditions are missing, but overall the experiment can be carried out by following the procedures (1 point)

❒ most steps are missing, no materials and conditions are included (0 points)

*Posttest.* The posttest served as a second criterion measure for assessing student performance. 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. The posttest items were directly aligned with the objectives of the instructional program. 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.

*Student attitude survey.* An 11-item attitude survey served as the criterion measure for assessing student attitudes and motivation toward the instruction. The items were four-choice Likert-type questions with the response choices being *strongly agree,* scored as (3), *agree* (2), *disagree* (1) and *strongly disagree* (0). These eight items dealt with topics such as: Did students like the program? Did they like conducting experiments? Were they confident in their ability to conduct experiments and write research reports as a result of the program? Internal reliability for the 8-item attitude survey, using Cronbach's alpha, was .72. The 3 remaining items were open-ended questions dealing with student likes, dislikes, and suggestions for improvement to the instructional program. Four additional Likert-type items were added to the 11-item student survey for subjects in the self-plus-teacher evaluation condition. These items asked about student attitudes toward self-evaluation.

Figure 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. There are at least 5 participants 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.

*Teacher attitude survey.* A 19-item attitude survey served as the criterion measure for assessing teacher attitudes. Nine items asked teachers about their delivery of the program, 6 Likert-type items assessed teacher attitudes toward various aspects of the instruction, and 4 open-ended questions asked teachers for suggestions on how to improve the instructional program. The survey for the self-plus-teacher evaluation group contained 3 additional items asking teachers about their attitudes toward student self-evaluation.

## Procedures

The researcher assigned the six teachers to one of the three treatment conditions (no-evaluation, teacher-evaluation, and self-plus-teacher evaluation). Each teacher taught two classes of students. In order to assign teachers to treatments, the researcher ranked all pairs of classes for each teacher from the highest achieving to the lowest achieving, based on the student 9th Grade Graduation Exam scores in mathematics and the Latvian language. The pairs of classes for each teacher were divided into high-achieving and low-achieving classes using a median split. Teachers with classes from each group were then randomly assigned to one of the three treatments. All experimental groups completed the *Learning Explorations* program with only the variations in evaluation conditions described below.

Students in the no-evaluation group conducted their experiments about learning and produced written reports. They were provided with the project rating scale before they began work on their experiments, but received no formal feedback from the teacher. They were not asked to evaluate their own work. Students in this group produced the initial experimental design diagrams before the start of the experiment, wrote the initial final report after the experiment, and could revise their work before submitting their final report to the teacher.

Students in the teacher-evaluation group conducted their experiments about learning and produced written reports. They received written teacher evaluation in the form of feedback on the initial version of their experimental design diagrams before the start of the experiment and on the initial versions of their final reports. The teachers used the project rating scale for evaluating student work. The teachers checked the most appropriate evaluative description for student research projects for each of the 15 criteria on the rating scale and wrote suggestions for improvement. Students were then able to revise their initial experimental design diagrams and initial final reports into final form by incorporating teacher suggestions into them.

Students in the self-plus-teacher evaluation group conducted their experiments about learning and produced written reports. They formally self-evaluated their work at the same two times (initial experimental design diagram and initial final report) during the instruction and used the same project rating scale as the teachers in the teacher-evaluation condition. In addition, the students received teacher feedback on their written reports. Once students had completed the self-evaluations of each product, they handed in to the teacher their initial experimental design diagram and their initial report with their self-evaluations of each. The teacher then wrote her evaluation on the project rating scale in the same manner as teachers in the teacher-evaluation condition. Students were then able to revise their work based on their own self-evaluation and on the teacher evaluation.

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 two treatments received additional instructions describing the evaluation procedures that they were expected to complete for their evaluation condition.

The teachers were told that the purpose of the research study was to investigate the effects of student self-evaluation and teacher evaluation on student performance. They were told that there were three 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 received no additional training on the use of the materials.

On average, teachers devoted 10 40-min class periods to teaching the program, with a range of 9 to 12 class periods. There were no significant differences among the treatment groups as to the time spent on the program. The no-evaluation group spent, on average, 11 class periods, while the teacher-evaluation and the self-plus-teacher evaluation groups spent, on average, 10 class periods. Two teachers taught the program as part of a high school psychology course, and the remaining four taught it during other classes. Students in all schools received a grade based on their posttest results and research project ratings. A day before the last class, the students took the course posttest. On the final day of the program, students submitted their

reports of their experiments and completed an attitude survey about the program.

The primary researcher, a Latvian doctoral student at an American university, traveled from the United States to Latvia to coordinate the study and observe its implementation. During the study, the researcher observed two class periods in each teacher's classroom to ensure that the instructional program was properly implemented and that the evaluation procedures for each treatment condition were closely followed. During classroom observations, the researcher wrote down cases when the teacher had difficulties with providing accurate explanations and responses to student questions or with conducting the instructional activities included in the program. After classroom observations, the researcher conducted an informal debriefing session with each teacher, asking them to share their overall feelings about the program.

After the experiment, the researcher also collected and reviewed the initial experimental design diagrams, the initial final research reports, and student and teacher evaluations of these reports. The researcher counted the number of cases where initial student reports had received lower than the maximum rating on each criterion of the project rating scale from the students and teachers. The researcher also examined the nature of student and teacher comments.

A diagram summarizing the experimental design is shown in Figure 2.

### Data Analysis

The data analysis for student performance was carried out as two separate one-way analyses of variance (ANOVAs), one for the student project scores and one for the posttest scores with three groups (no-evaluation, teacher-evaluation, and self-plus-teacher evaluation) in each analysis. Student attitude data were analyzed using a 3 (Treatments) × 8 (Survey Items) multivariate analysis of variance (MANOVA) for the mean scores on the survey items. To control for Type I error across univariate follow-up tests for the eight items on the student survey, Bonferroni correction was used setting alpha level at .006 (.05/8 items). The frequency of constructed responses to the three open-ended questions on the attitude survey was computed and is reported later in this paper. Mean scores on the Likert-type items and constructed response on the open-ended items for the teacher survey were also analyzed.

## RESULTS

Results are discussed below by achievement, student attitudes, teacher attitudes, student and teacher ratings of initial research reports, and classroom observations.

### Achievement

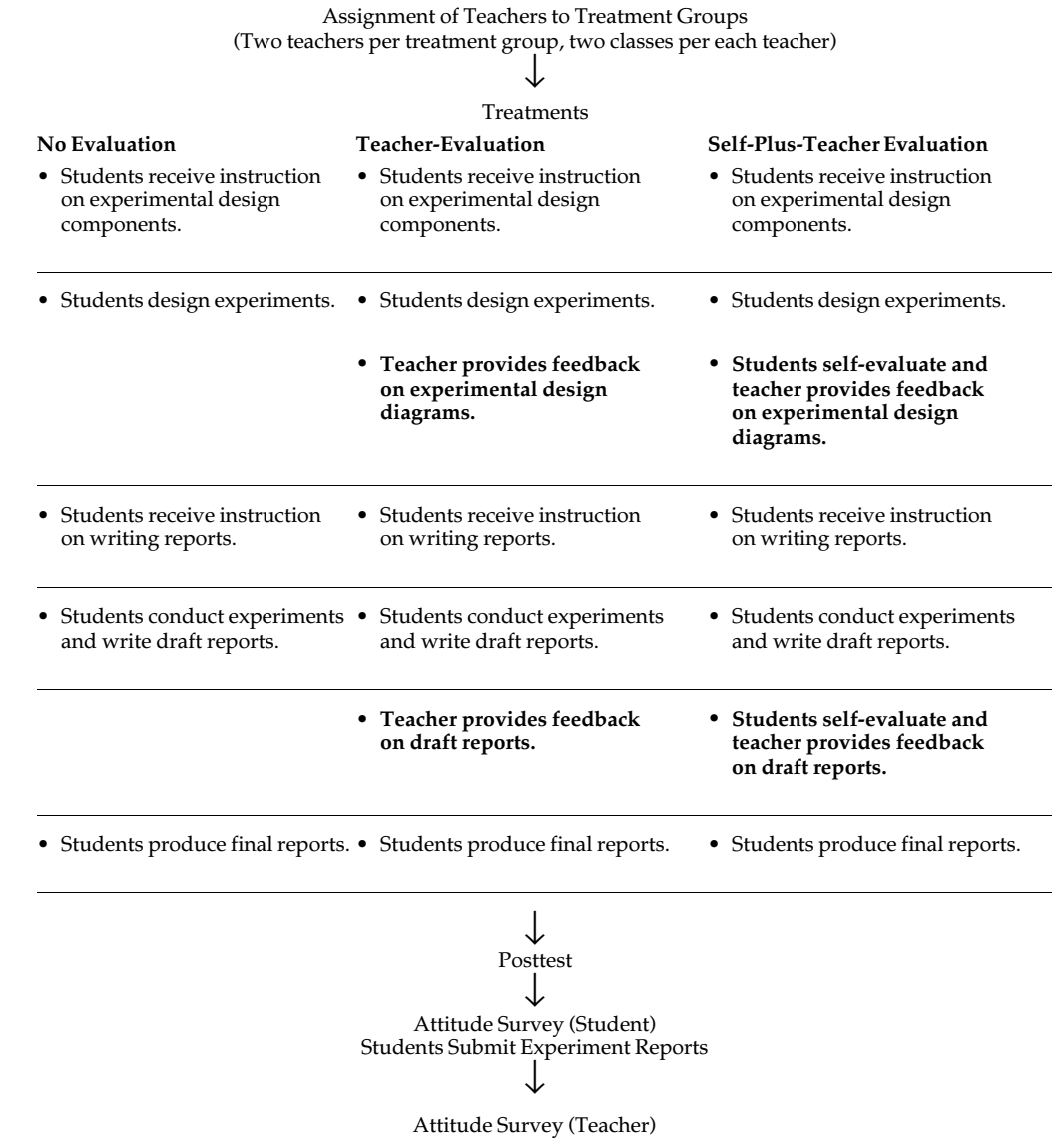
The mean scores and standard deviations for both the project reports and the posttest are shown in Table 1.

*Ratings of Student Research Reports.* The table reveals that the mean scores for the project reports were 14.87 (50%) for the no-evaluation group, 17.49 (58%) for the teacher-evaluation group, and 18.10 (60%) for the self-plus-teacher evaluation group.

The ANOVA conducted on the project report scores yielded a significant overall difference, $F(2, 186) = 5.70$, $p < .01$. Because the overall $F$ value was significant, follow-up Fisher's least significant difference (LSD) tests were performed to determine whether significant differences occurred between the mean scores for each pair of treatments. These tests revealed that both the teacher-evaluation group and the self-plus-teacher evaluation group had significantly higher scores at the $p < .01$ level on their project reports than the no-evaluation group. The difference in the project report scores between the teacher-evaluation group and the self-plus-teacher evaluation group was not statistically significant.

*Posttest.* The mean posttest scores were 21.87 (73%) for the no-evaluation group, 23.96 (80%) for the teacher-evaluation group, and 22.78

Figure 2 ☐ Summary of Experimental Design




(76%) for the self-plus-teacher evaluation group. The ANOVA conducted on the posttest scores yielded a significant overall difference, $F$ (2, 186) = 4.11, $p < .05$. Follow-up LSD tests revealed that the teacher-evaluation group scored significantly higher at the $p < .01$ level than the no-evaluation group. The differences in the posttest mean scores were not statistically significant between the no-evaluation group and the self-plus-teacher evaluation group or between the teacher-evaluation group and the self-plus-teacher evaluation group.

### Student Attitudes

The mean attitude scores by treatment for the student responses to the eight statements on the four-point Likert-type attitude survey administered after completion of the instructional

Table 1 ☐ Mean Project Report and Posttest Scores by Treatment Group

| | Treatment | | |
|---|---|---|---|
| Measure | No Evaluation (n = 62) | Teacher Evaluation (n = 69) | Self + Teacher Evaluation (n = 58) |
| Project Reports | | | |
| *M* | 14.87 | 17.49 | 18.10 |
| *SD* | (5.65) | (5.83) | (5.39) |
| Posttest | | | |
| *M* | 21.87 | 23.96 | 22.78 |
| *SD* | (4.17) | (3.71) | (4.68) |

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

program are shown in Table 2. Responses were scored on a four-point scale from 3 for the *most positive* response to 0 for the *most negative* response.

The overall mean score across the eight Student Attitude Survey items was 1.87, a moderately favorable rating indicating general agreement with positive statements about the instructional program. The three highest-rated statements on the survey were "I now understand how to conduct experiments" ($M$ = 2.11), "The program helped me learn how to conduct experiments" ($M$ = 2.07), and "I am satisfied with my experiment and report" ($M$ = 2.07). The two lowest-rated statements were "I liked planning and conducting experiments" ($M$ = 1.54) and "The information was easy to understand" ($M$ = 1.70).

The data in Table 2 were analyzed using a 3 (Treatment) × 8 (Survey Items) MANOVA to test for significant differences. The overall means were significantly different for the three treatment groups, Wilks's $\Lambda = .723$, $F$ (16, 322) = 3.54, $p < .001$. Follow-up univariate ANOVAs were conducted on each of the eight items. Using Bonferroni adjustment, each ANOVA was tested at .006 level (.05/8 items). The analysis revealed significant attitude differences between the treatment groups on three of the items.

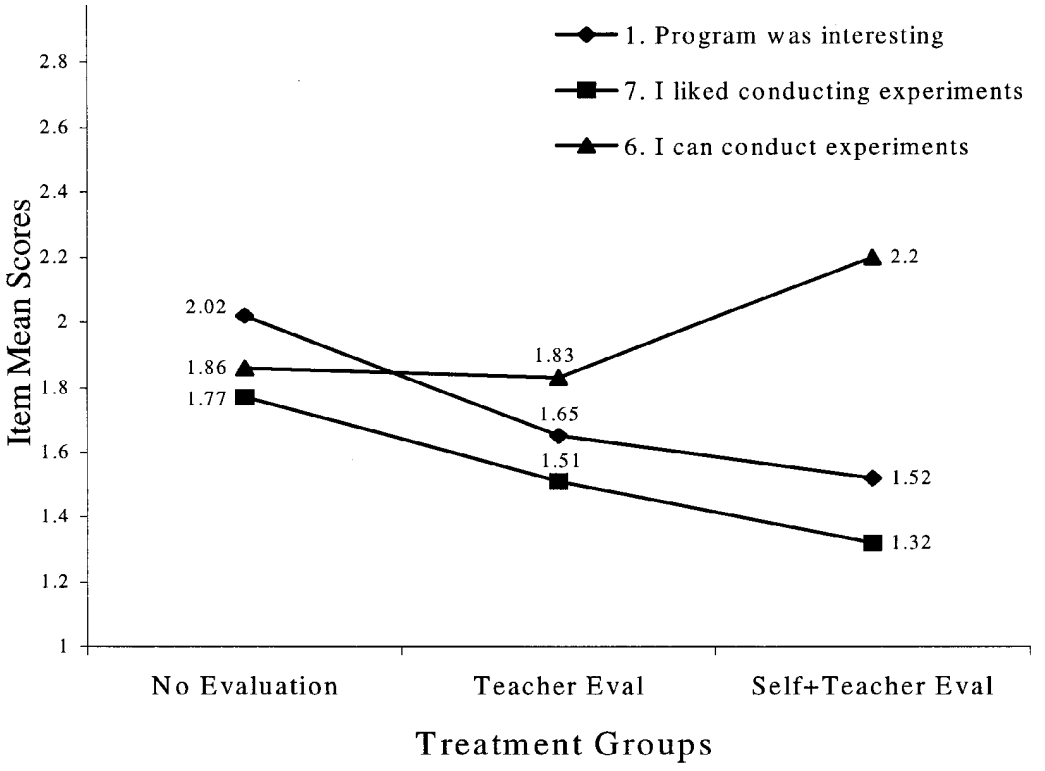
Posthoc LSD tests setting alpha level at .006 were performed on the three significant items. As shown in Figure 3, students in the no-evaluation group had significantly more favorable

Table 2 ☐ Mean Ratings on Student Attitude Survey

| | Treatment | | | | |
|---|---|---|---|---|---|
| Item | No Evaluation | Teacher Evaluation | Self + Teacher Evaluation | *F* | *p* |
| 1. The program was interesting. | 2.02 | 1.65 | 1.52 | 10.89 | < .006* |
| 2. It was easy to learn from the . program | 1.73 | 1.68 | 1.92 | 2.09 | ns |
| 3. The information was easy to understand. | 1.84 | 1.51 | 1.78 | 3.54 | ns |
| 4. The program helped me learn how to conduct experiments. | 2.20 | 2.00 | 2.00 | 1.29 | ns |
| 5. I now understand how to conduct experiments. | 2.04 | 2.14 | 2.14 | .55 | ns |
| 6. I can now independently plan and conduct experiments. | 1.86 | 1.83 | 2.20 | 5.77 | .006 |
| 7. I liked planning and conducting experiments. | 1.77 | 1.51 | 1.32 | 5.32 | .006 |
| 8. I am satisfied with my experiment and report. | 2.18 | 2.08 | 1.94 | 2.20 | ns |
| Overall means | 1.96 | 1.80 | 1.85 | 3.54 | < .001 |

* .006 was the probability level at which differences were tested for significance using the Bonferroni correction (.05/8 items).

Figure 3 ☐ Mean Differences by Treatment for Statistically Significant Student Attitude Items




scores on the item, "The program was interesting," than students in the teacher-evaluation and self-plus-teacher evaluation groups ($M$ = 2.02 for no-evaluation, $M$ = 1.65 for teacher-evaluation, and $M$ = 1.52 for self-plus-teacher evaluation). In addition, students in the no-evaluation group had significantly more favorable scores on the item, "I liked planning and conducting experiments," than students in the self-plus-teacher evaluation group ($M$ = 1.77 for the no-evaluation, and $M$ = 1.32 for the self-plus-teacher evaluation). On the other hand, students in the self-plus-teacher evaluation group had significantly more favorable scores on the item, "I can now independently plan and conduct my own experiments," than students in the no-evaluation group and in the teacher-evaluation group ($M$ = 2.20 for the self-plus-teacher evaluation, $M$ = 1.86 for the no-evaluation, and $M$ = 1.83 for the teacher-evaluation).

The mean score for the four additional items on the student survey for the self-plus-teacher evaluation group was 1.90, indicating a moderately favorable attitude toward self-evaluation. The two highest-rated items were "My self-assessment was honest" ($M$ = 2.20) and "My self-assessment was important to the teacher" ($M$ = 2.12). The two lowest-rated items were "I liked assessing my own work" ($M$ = 1.52), and "When I have to assess my own work, I know what to improve" ($M$ = 1.77).

Summary of the open-ended responses on the student attitude survey indicated that what students liked most about the program was participating in and conducting their own experiments, a response given by a total of 53 of the 189 students (28%) across the three groups. The second most common response to what students liked most was the fact that information in the program was well presented and comprehensive, a response indicated by 41 students (22%), and third was that there were interesting and

useful experiment examples (35 students, 19%). When asked what they liked least about the program, 53 students (28%) indicated the use of complex language and difficult terms, 36 (19%) reported that they disliked too much reading and redundant examples, and 25 (13%) responded that they liked everything. When asked for suggestions on how to improve the program, 52 students (28%) suggested shortening and simplifying text, 22 (12%) recommended including more practical experiments and exercises to be conducted in class, and 12 (6%) indicated that more class time should be allowed for delivery of the program. The pattern of student comments did not differ appreciably across the treatment groups.

### Teacher Attitudes

The six Likert-type items on the teacher survey were scored on a three-point scale from 2 for the *most positive* response to 0 for the *least positive* response. The overall mean rating for the six teachers on the survey was 1.53, a favorable rating indicating agreement with positive statements about the instructional program. The highest-rated item was "The material develops student skills in designing and conducting their own experiments well" ($M = 1.83$). All teachers said that the lesson procedures in the teacher guide were clear, that they followed the teacher guide closely when working with the students, and that they would use the material again in their work with students.

The two teachers in the self-plus-teacher evaluation group thought that the program developed student understanding of the basic experimental design concepts in psychology to a greater extent than did the teachers in the other two treatment groups. When asked about their attitudes toward student self-evaluation, the two teachers in the self-plus-teacher evaluation group felt strongly that self-evaluation helped students produce higher quality experiment reports and that they would use self-evaluation in their teaching in the future.

### Student and Teacher Ratings of Initial Research Projects

Student and teacher ratings of initial experimental design diagrams and initial research reports were also examined. Examination revealed that students in the self-plus-teacher evaluation group tended to rate their initial reports higher than did their teachers. Students in the self-plus-teacher evaluation group assigned lower than maximum ratings of 2 on an average of only 3 of the 15 criteria on the initial report, while teachers in both groups with the teacher-evaluation component assigned lower than maximum ratings of 2 on an average of 6 of the 15 criteria. In addition, student ratings of their reports did not contain additional comments with ideas for improvement, while teacher evaluations contained several such comments. Teacher comments, for the most part, provided students with knowledge of results, such as "You did not describe how the treatment groups were formed" or "You did not list all constants".

### Classroom Observations

Classroom observations by the experimenter revealed that on several occasions teachers had difficulty delivering the program optimally because of their lack of sufficient content knowledge. In addition, it was difficult for the teachers to properly carry out classroom demonstration experiments described in the teacher guide because they had not observed experiments themselves. During informal debriefing sessions with the experimenter after the classroom observations, several teachers noted that rating the student project reports was very time consuming and that it was difficult for them to fit this task into their busy schedules.

## DISCUSSION

This study examined the effects of teacher evaluation and the combination of teacher evaluation and student self-evaluation on student performance and attitudes. Students in the teacher-evaluation and the self-plus-teacher

evaluation conditions received significantly higher ratings on their research projects than did students in the no-evaluation condition. In addition, students in the teacher-evaluation condition scored significantly higher on the posttest than the other two groups. Students in the no-evaluation group had more positive attitudes toward the program than did students in the teacher-evaluation and the self-plus-teacher evaluation group, and they reported that they enjoyed conducting experiments more than did students in the self-plus-teacher evaluation group. However, students in the self-plus-teacher evaluation group had greater confidence about their ability to independently conduct experiments in the future than the other two groups.

The first research question asked if teacher evaluation would have a positive effect on student performance. Students who received teacher evaluation, with or without self-evaluation, produced higher quality reports of their experiments than students who did not. Teachers most likely had greater knowledge than students about how to conduct experiments and write reports. This knowledge would help the teachers to provide higher quality feedback to students than the students could generate on their own. The better feedback, in turn, would help the students learn what they needed to improve and how to improve it. Analysis of the differences between student and teacher ratings of the initial research projects revealed that teachers assigned lower than maximum ratings on twice as many criteria on the project rating scale as did students. In addition, teachers also provided students with additional comments suggesting improvements to their reports, while student self-evaluations contained almost no comments. Thus, the better student performance on research reports under teacher-evaluation conditions would appear to be due to better evaluation and feedback provided by the teachers.

The significantly higher scores on the posttest for the students in the teacher-evaluation group over those in the no-evaluation group may have been due, at least in part, to information provided in the teacher evaluation/feedback that was directly relevant to the program content assessed on the posttest. This information would have been covered in the direct instruction for both groups, but may often have been reemphasized in the teacher evaluation/feedback to students who did not apply it well to their experimental designs and draft reports, thus enabling them to learn it better prior to the posttest. In addition, reading student initial reports and providing feedback might have enabled teachers to gain information on student performance during instruction. Teachers could then use this information to improve their instruction by reteaching certain concepts as necessary, thereby helping students to improve both student posttest scores and the quality of their research projects.

The second research question asked whether the combination of teacher evaluation and student self-evaluation would have a different effect on student performance than teacher evaluation alone. There was no strong evidence from the study that self-evaluation when used in combination with teacher evaluation produced an improvement in student performance over using teacher evaluation alone. The fact that the addition of self-evaluation did not significantly improve student performance could be attributed to the greater evaluation expertise of the teachers described above. Most likely, teacher evaluations were more complete and accurate than student self-evaluations, and student self-evaluations therefore did not add significantly to the quality of student projects. Another reason for the lack of improvement in student performance could be that students were not familiar with self-evaluation as an evaluation strategy because it is rarely used in Latvian schools. Providing students with practice in self-evaluation, such as applying the project rating scale to research report examples, might help students gain a better understanding of the use of evaluation criteria and become more accurate at evaluating their own work.

The mean scores on the student projects were lower than on the posttest, with overall means of 16.8 out of 30 possible for the project reports and 22.9 out of 30 possible for the posttest. Producing the research projects and reports was a more difficult and advanced task than performing well on the posttest. The posttest required only

recognition and recall of concepts and application of knowledge to scenarios provided by the experimenter. Planning an experiment and producing a research report, on the other hand, required application of knowledge to generate new products in the form of an experimental design and a written report. In addition, designing an experiment and producing a written research report required more self-direction by the students than responding to the posttest. Students had an especially hard time in their experiment reports coming up with a precise description of experimental procedures and writing conclusions. Several students had difficulties managing their experimental subjects and properly implementing their intended experimental procedures.

The third research question related to the effects of the experimental treatments on student attitudes. The attitude results revealed that students in the no-evaluation condition were more interested in the instructional program than those in the other two experimental treatments and that they liked the program significantly better than those in the self-plus-teacher evaluation condition. These results may be attributable to the fact that the program overall was easier for students in the no-evaluation condition. They did not have to formally evaluate their work and they did not receive any feedback either from the teacher or from their own self-evaluations regarding changes needed to improve their projects. The finding that students have more positive attitudes toward treatments that are easier, but less effective instructionally, is consistent with the results of studies by other researchers (Hannafin & Sullivan, 1996; Schnackenberg, Sullivan, Leader & Jones, 1998).

That students in the self-plus-teacher evaluation group reported significantly greater confidence in their ability to independently conduct experiments in the future than their counterparts in the other two groups may have been related to their involvement in the self-evaluation process. Formal self-evaluation of their projects may have caused these students to think that they better knew the criteria for designing and reporting a research project. It is possible that they felt more in control of their learning than students in the other two groups who did not evaluate their own work. Their greater confidence in their ability supports the findings of other researchers that students who self-monitor and self-evaluate their progress have higher self-efficacy perceptions than those who do not (Schunk, 1989; Zimmerman & Kitsantas, 1999).

The three highest-rated items on the student attitude survey all dealt with students' learning how to conduct experiments and with their satisfaction with their experiments and reports. These items dealt quite directly with attitudes toward desired outcomes of the program, that is, conducting and reporting experiments, rather than with process variables such as whether the program was interesting, easy-to-understand, or easy to learn from. It is encouraging that students generally agreed that they acquired these desirable outcomes irrespective of their treatment group.

The highest-rated item by the teachers, "The program develops student skills in designing and conducting their own experiments well," is quite consistent with highly rated items by the students indicating that the program was effective in helping them to learn to conduct and report experiments. In addition, despite the lack of strong evidence that self-evaluation improved student performance, teachers in the self-plus-teacher evaluation group showed support for self-evaluation by indicating that it helped students produce higher quality reports and that they would use it in their teaching in the future.

This study revealed that providing students with teacher evaluation and feedback in the formative stages of student work and having students incorporate the feedback into their final products improves student performance. Yet the teachers considered these strategies to be too time-consuming. Development of formative evaluation strategies that are less time-consuming may be necessary to ensure their use by teachers. One option may be to explore further the effects of self-evaluation on student performance in order to provide students with timely formative feedback without overburdening the teacher. Another option may be to have teachers provide group feedback to the students regarding the common strengths and weaknesses of their projects during the formative stages of their work. Emphasizing sound instruction that

is aligned with objectives and providing students with multiple guided-practice opportunities during instruction might also reduce the need for detailed teacher feedback on student work during the evaluation stage.

Classroom observations revealed that, even though teachers were provided with detailed instructional procedures in the teacher guide, they did not do a particularly good job of teaching the instructional content that, in general, was relatively unfamiliar to them. The difficulties that the teachers had delivering the instructional program could be attributed to the fact that instructional methods and assessment strategies included in the program were relatively new to the Latvian teachers. The program included project-based teaching requiring teachers to facilitate student research projects. In addition, the *Learning Explorations* program included performance assessment as the main indicator of student achievement, which contrasts with more traditional assessments in Latvian schools that focus primarily on factual, conceptual and procedural learning. Providing teachers with pretraining on the content and use of the instructional program is one option that might enable them to deliver instructional programs containing relatively unfamiliar content and novel instructional approaches more effectively.

Limitations on the number of classes available for this study precluded use of a student self-evaluation-only treatment that would have permitted analysis of the unique contribution of self-evaluation to student performance and attitudes. Adding a student self-evaluation group to the design of a future study would address this issue and permit a more complete analysis of the effects of self-evaluation and teacher evaluation. Discussing with students the importance of self-evaluation and providing them with instruction on the use of an evaluation instrument such as the project rating scale from this study might also help them improve their own self-evaluations. Future research that investigates potentially productive ways to incorporate teacher evaluation and student self-evaluation into classroom instruction should help us understand the most effective strategies for using classroom evaluation to improve student learning. □

Zane Olina [olina@coe.fsu.edu] is an assistant professor of Instructional Systems at Florida State University, Tallahassee, and Howard J. Sullivan [sully@asu.edu] is a professor of Educational Technology at Arizona State University, Tempe.

This study was conducted while Dr. Olina was a student at Arizona State.

## REFERENCES

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

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.

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.

Crooks, T.J. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research, 58,* 438–481.

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.

Hannafin, R.D., & Sullivan, H.J. (1996). Preferences and learner control over amount of instruction. *Journal of Educational Psychology, 88,* 162–173.

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

Latvian Ministry of Education and Science Center for Curriculum Development and Examination. (1998). *National Standards of Compulsory Education.* Lielvarde, Latvia: Lielvards Ltd.

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

Natriello, G. (1987). The impact of evaluation processes on students. *Educational Psychologist, 22,* 155–175.

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.

Schnackenberg, H.L., Sullivan, H.J., Leader, L.F., & Jones, E.E.K. (1998). Learner preferences and achievement under differing amounts of learner practice. *Educational Technology Research and Development, 46*(2), 5–15.

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.

Scriven, M. (1967). The methodology of evaluation. In R.W. Tyler, R. Gagné, & M. Scriven (Eds.), *Perspectives of curriculum evaluation: Vol. 1. AERA monograph series on curriculum evaluation* (pp. 39–83). Chicago: Rand McNally.

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

Soros Foundation-Latvia. (2001, December). *A passport to social cohesion and economic prosperity: Report on education in Latvia in 2000* [Online]. Available URL: http://www.politika.lv/polit_real/files/lv/Izgp_en1.pdf

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.

Story, N.O., & Sullivan, H.J. (1986). Factors that influence continuing motivation. *Journal of Educational Research, 80,* 86–92.

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.

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