Education Tech Research Dev (2008) 56:181–201
DOI 10.1007/s11423-006-9012-7

DEVELOPMENT ARTICLE

Teacher, self and peer evaluation of lesson plans written by preservice teachers

Gamze Ozogul · Zane Olina · Howard Sullivan

Published online: 17 October 2006

Abstract The study investigated the effects of three types of evaluation on preservice teachers' performance, knowledge and attitudes related to writing lesson plans that incorporate technology. Subjects were randomly assigned to one of the three treatment conditions: teacher-evaluation, self-evaluation or peer-evaluation. All groups completed three class periods of instruction on writing lesson plans, then each subject submitted his/her draft lesson plan. The drafts were evaluated by assigned evaluators (teacher, self or peer), who provided scores and written feedback on a 15-item rubric. Students then revised their lesson plans into final form. All three treatment groups improved their lesson plans significantly from draft version to final version, with the teacher-evaluation group showing significantly greater improvement and writing significantly better final lesson plans than each of the other two groups. Teacher-evaluation and self-evaluation groups had significantly higher scores on a knowledge-based posttest than the peer-evaluation group. Several suggestions are discussed for making further improvements in the self-evaluation and peer-evaluation processes.

G. Ozogul (✉)
Psychology in Education Division, College of Education, Arizona State University, Tempe 85287-0611 AZ, USA
e-mail: Gamze.Ozogul@asu.edu

Z. Olina
Department of Educational Psychology and Learning Systems, Florida State University, 307 Stone Building, Tallahassee 32306-4453 FL, USA
e-mail: olina@coe.fsu.edu

H. Sullivan
Psychology in Education Division, College of Education, Arizona State University, Tempe 85287-0611 AZ, USA
e-mail: Sully@asu.edu

Keywords Evaluation · Teacher · Self · Peer · Preservice · Lesson plan · Research

Introduction

Scriven (1967) defined formative evaluation as evaluation conducted for the purposes of improving educational programs that are still in the development process. Formative evaluation procedures are frequently used to identify potential improvements during pilot testing or field testing of new educational programs. These improvements are then incorporated into the program before it is made available for general users.

Researchers have extended the concept of formative evaluation beyond the evaluation of instructional programs in recent years by applying it to the evaluation of student work that is still under development (Fallows & Chandramohan, 2001). The typical process here is for students to use given guidelines or specifications to develop a product (e.g., a research report, a business plan) in draft or near-final form. The student products are then evaluated, often using the same guidelines, and feedback based on the evaluation is provided to each student. The students then use this feedback to revise their products into final form. Such formative evaluation has been widely discussed as a strategy with great potential to enhance student learning and increase learner motivation (Black & William, 1998; Sadler, 1989; Wolf, Bixby, Glenn, & Gardner, 1991). 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.

Three types of individuals have served as evaluators in formative evaluation of student work: the teacher, the student him/herself, and student peers. There is a considerable amount of research evidence on the effects of formative evaluation by teachers on student learning, less on self-evaluation and much less on peer-evaluation. This study was conducted to provide a direct comparison of the effects of the three evaluation conditions on both student performance and knowledge.

Research on the effects of teacher evaluation on student performance when used for formative evaluation purposes has generally yielded positive results (Black & William, 1998). Olina and Sullivan (2002) found that high school students who received formative teacher-evaluation with or without self-evaluation produced higher quality reports of their experiments and scored significantly higher on a knowledge-based posttest than students who did not receive formative evaluation by the teacher. Elawar and Corno (1985) reported that sixth-grade students' performance and attitudes toward mathematics were improved when teachers provided constructive written feedback on their homework on a weekly basis.

A problem associated with formative evaluation of student work by the teacher is the amount of teacher time required to do the evaluation and to provide useful feedback to the students. Both student self-evaluation and peer-evaluation shift much of the responsibility of formative evaluation to the

students, and in doing so, may considerably reduce the teacher's workload in the formative evaluation process (Ballantyne, Hughes, & Mylonas, 2002; Fallows & Chandramohan, 2001).

More importantly, student self-evaluation and peer-evaluation are believed to have important learning benefits for students in addition to potentially reducing teacher workload. Support for use of student self-evaluation in instruction comes from several areas of research, including metacognition (Bransford, Brown, & Cocking, 1999; Flavell, 1976; Winne & Hadwin, 1998) and classroom assessment (Gipps, 1994; Sadler, 1989; Stiggins, 2001). From a metacognition perspective, formal engagement in appraisal of their work prompts students to monitor their levels of mastery and understanding and to detect shortcomings. From a classroom assessment perspective, self-evaluation enables students to take a more active role in their own learning and may help them gain an important long-term skill that improves the quality of their work without relying heavily on others (Davies, 2002; Fallows & Chandramohan, 2001; Wiggins, 1998; Wolf et al., 1991). According to Chi, Bassok, Lewis, Reimann, and Glaser, (1989), good students engage in self-explanations and justification of their actions to a greater extent than do poor students. Engagement in both formal self-evaluation and peer-evaluation is likely to serve as a catalyst for such self-explanations.

Researchers have suggested that self-evaluation has the potential to improve student performance (Davies, 2002; Fallows & Chandramohan, 2001). Student self-evaluation is thought to develop critical thinking skills and enable students to take a more active role in their own learning. Critically reflecting on their own work may help students gain an important long-term skill that improves the quality of their work without relying heavily on others (Gipps, 1994; Wiggins, 1998; Wolf et al. 1991).

However, studies on the effects of self-evaluation on student learning have yielded mixed results. On the positive side, Kitsantas and Baylor (2001) found that preservice teachers who used a self-regulatory tool to evaluate themselves scored higher on their posttest than did students who did not use self-evaluation. Similarly, Fontana and Fernandes (1994) reported that primary school students who were trained in self-evaluation and evaluated themselves in math scored significantly higher on a mathematics test than a control group that did not use self-evaluation. In contrast, Andrade and Boulay (2003) reported no effect of student self-evaluation on the performance of seventh-grade and eighth-grade students on written essays. Also, in two studies conducted with high school students, Olina and Sullivan (2002, 2004) found that student formative self-evaluation of their own draft research reports did not result in improvements on the final versions of their research reports.

Peer-evaluation has the potential to be beneficial to both the assessor and the assessed. Peer-evaluation can increase the amount of time students spend on task, their level of engagement, and the amount of practice they receive. It can also result in a greater sense of accountability and responsibility for their work (Liu, Lin & Yuan, 2002; Topping, Smith, Swanson & Elliot, 2000). Furthermore, receiving a greater amount of feedback on their own work and comparing their work with the work of their peers may result in greater

metacognitive awareness and contribute to development of self-evaluation skills (Topping et al., 2000).

Much of evidence on the effects of peer-evaluation comes from student reports of its effects, rather than from measurement of student learning in a peer-evaluation context. However, several researchers have cited positive effects based on student reports. Brindley and Scoffield (1998) reported that business undergraduate students who used peer-evaluation on instructional modules noted that it increased their personal motivation and their understanding of the instructional content. Davies (2000) found that college students reported receiving significant benefits from marking the work of their peers, with most students indicating that they had a deeper level of understanding of the learning content after the peer-evaluation process. Smith, Cooper and Lancaster (2002) reported that peer-evaluation increased university undergraduate students' awareness of the evaluation process and their use of the scoring criteria in their own work.

When student attitudes toward teacher, self and peer evaluation are compared, students seem to prefer teacher-evaluation over the other two strategies. Zhang (1995) explored college students' preferences among teacher, self and peer evaluators in a formative-evaluation process. He found that a majority of his subjects preferred teacher feedback over the two other choices, and a majority chose peer-evaluation over self-evaluation when comparing only those two forms. Similarly, Olina and Sullivan (2004) found that students liked teacher-evaluation better than self-evaluation.

An important issue when considering use of self and peer assessment in the classroom is how well students can appraise their work, or at least to what extent their evaluations are consistent with that of the teacher. A number of researchers have compared student-generated ratings with those generated by the teacher. Falchikov and Boud (1989), in their analysis of quantitative self-assessment studies in higher education, found no conclusive evidence regarding consistency of student and teacher ratings, but suggested that there was a tendency by the students in most studies to overrate or underrate their own performance. Liu, Lin and Yuan (2002) examined the correspondence of ratings across different combinations of evaluators in the formative evaluation process: peer-teacher, self-teacher and self-peer. Their results showed that the scores assigned by self-evaluators were significantly higher than those assigned by peer evaluators, and peer-evaluator scores were significantly higher than those assigned by teacher evaluators.

Student perceptions of their peer evaluators and their own role as peer assessor appear to affect the type of feedback they provide and the importance they assign to peer feedback. Researchers report that students have more favorable attitudes toward evaluating peers' work when the peer evaluator is kept anonymous to the student being evaluated. Brindley and Scoffield (1998) found that when the student being evaluated knew who his/her peer evaluator was, the peer evaluator often did not assign a low grade even if it were appropriate. Falchikov (1986) reported a tendency for known peer evaluators to grade their peers' work higher than self-graders. In addition, Miller and Ng (1994) reported that student attitudes were negative toward peer-evaluation when a known peer provided evaluation and feedback.

The authors attributed this effect to the peer evaluator not wanting to make unfavorable judgments about the person s/he evaluates. In line with these findings, Cheng and Warren (1997) found that undergraduate college students did not complain of subjectivity and unfairness when the student and peer evaluator were anonymous to each other. Similarly, Davies (2000) reported that classmates show a better respect for judging the quality of work when they are anonymous in the peer-feedback process.

Overall, teacher-evaluation is the most common form of classroom evaluation and studies of its effects as a formative evaluation procedure have generally yielded positive results with regard to student learning. Researchers have cited several potential benefits of both self- and peer-evaluation, but the effects of these two evaluation procedures on student learning are much less clear-cut than the effects of formative evaluation by teachers. Self-evaluation has yielded mixed learning results across studies, and most studies of peer-evaluation have used student and/or researcher attitudes or reports, rather than direct measures of student learning, as outcome measures.

The purpose of the present study was to provide a direct comparison of effects of teacher-evaluation, self-evaluation, and peer-evaluation on the achievement and attitudes of preservice teachers. Two achievement measures were used: (1) a performance measure for which the preservice teachers developed a lesson plan that integrated use of technology into the lesson plan and (2) a knowledge measure in the form of a 15-item posttest that assessed preservice teachers' acquisition of learning content presented during the instructional phase of the study. In addition to the achievement and attitude information, data were also collected on the teacher, self and peer ratings of the lesson plans in their initial form; the nature of feedback provided by each type of evaluator, and the judgments of the course instructors about the three types of evaluation.

The study was conducted in an instructional design course for preservice teachers that focused on integrating technology into the classroom. The performance task required the students to develop a lesson plan that was based on systematic instructional design principles and incorporated the use of technology. The focus of the course and the lesson plan unit was consistent with the recent emphasis on training teachers to integrate technology into their instruction (Doering, Hughes & Hoffman, 2003; Roberts & Hsu, 2000). According to Reiser (1994), teachers are more likely to employ an effective systems approach in their future instructional planning when they receive appropriate training. Further, self- and peer-evaluation experience may be useful to future teachers in their teaching careers.

Method

Subjects

The subjects were 101 preservice teachers in the undergraduate teacher education program at Arizona State University. Their average age was 23.

About 36% of the subjects were male and 64% were female. All subjects were juniors preparing to be secondary school teachers. They had no prior experience with incorporating technology into their lesson plans or into instruction. They were enrolled in one of six sections of a required introductory instructional design course focusing on integrating technology in the classroom. The course was taught over an eight-week period. The average section size was 17 students.

Three instructors taught two sections each of the course. All three were doctoral students in either Educational Technology or in Science Education. The three instructors had considerable knowledge of the course content, as well as previous experience teaching the course.

Research design

The research design for the lesson-plan performance measure was a 3 treatment (teacher-evaluation, peer-evaluation, self-evaluation) $\times$ 2 lesson-plan version (draft lesson plans and final lesson plans) design. This design permitted analysis and comparison of student performance on the lesson plans both prior to formative evaluation of the plans and after formative evaluation and revision of them. Additional quantitative data were collected using a 15-item posttest measure of student knowledge administrated after submission of final lesson plans.

The quantitative data described above were complemented by several sources of qualitative data. These sources included student attitudes and perceptions collected in a student attitude survey administered at the end of the study, teacher attitudes and observations collected in individual interviews after the study, and the nature of the feedback comments provided by each type of evaluator during formative evaluation of the lesson plans. The combination of quantitative and qualitative measures was designed to provide a broad base of descriptive and comparative data regarding student performance and knowledge, student and instructor attitudes and judgments about the treatments, and the nature of various formative-evaluation comments made by students and instructors.

Procedures

The instructors were trained in the procedures for the study by the primary researcher. During the training the instructors were provided with a detailed syllabus for the lessons and an instructor guide, which they were directed to follow closely. The researcher then randomly assigned the six sections to three treatment conditions (teacher-evaluation, self-evaluation and peer-evaluation), so that two sections were assigned to each condition and each instructor taught two different conditions.

All groups completed the same instruction for the first three weeks of the class. During this time, subjects were provided with information and practice on the parts of a lesson plan, including content on the integration of

appropriate technologies and two video cases that showed examples of teachers effectively integrating technology into their classroom instruction. In addition, during week 2 all students were provided with a 15-item lesson-plan rubric, described below in the criterion measures section, to use in developing and evaluating the technology-integrated plans. Students worked on their lesson plans weekly by adding parts of the plan covered in the class. All students used an online portfolio tool for building and storing their plans.

During week 4, procedures varied across the three treatment groups. Subjects in the teacher-evaluation condition received teacher-evaluation and feedback on their individual lesson plans. Those in the self-evaluation condition provided their own evaluation and feedback on their own lesson plan during class time. Individuals in the peer-evaluation group provided evaluation and feedback on the lesson plan of one anonymous peer during class time.

Subjects in the teacher-evaluation condition were told during the week 4 class that their draft plan would be evaluated by the teacher. They were also told that they should revise the plan based on the teacher feedback and submit the revised lesson plan at the beginning of the week 5 class. The teachers collected the initial lesson plans and used the standard 15-item rubric described later in this section to evaluate each lesson plan.

The teacher-evaluators were instructed to rate each rubric item from 0–2 and they were also told to provide written feedback on at least four of the items on each student's lesson plan. Teachers informed the students during the class that the lesson-plan evaluations would be available to them later in the same day of the class. They offered two options for returning the lesson-plan evaluation, either to pick it up from the office or to receive it via e-mail.

Subjects in the self-evaluation condition were told during the week 4 class that they would evaluate their own lesson plan. They were also told that they should revise the plan based on their own written feedback from their evaluation and submit the revised lesson plan at the beginning of the week 5 class. The teacher told the students to formally evaluate their own plan during the class time by using the 15-item rubric. The students were instructed to rate each item on the lesson-plan rubric from 0–2 and to provide written feedback on at least four items.

Subjects under peer-evaluation were told during the week 4 class that their initial lesson plan would be evaluated by one of their classmates. They were also told that they should revise it based on the peer feedback and submit the revised plan in the week 5 class. The teachers collected the draft lesson plans during the week 4 class, covered the author name on them, and gave one lesson plan to each student to evaluate. The study was designed so that the peer feedback would be anonymous, and considerable care was taken before and during the treatment to preserve anonymity. Each student evaluated the lesson plan of one anonymous peer during class time by using the standard 15-item evaluation rubric. Peers were instructed to rate each item from 0–2 and to provide written feedback on at least four items. After the peers had evaluated the plans, the teacher collected them and returned them to their original authors during class.

The minimum of four items on the 15-item rubric was decided upon by the researchers because many of the students in the pilot test conducted prior to this study made very few feedback comments. Four comments on the lesson plan seemed to the researchers to be a feasible minimum number to provide reasonable guidance to the lesson-plan writers.

In summary, the procedures for evaluating the draft lesson plans were identical for all three types of evaluators. Only the type of evaluator varied across these treatment groups.

Materials

The first five weeks of this eight-week course were taught by the three teachers using a printed instructor guide and student handouts designed by the researcher for the purposes of the study. The instructor guide included detailed procedures for lesson delivery, master handouts for students, and information for using the handouts. The instructor guide also included the procedures for each treatment condition. A pilot test of the instruction was conducted with three other classes in the same course before the intervention, and the materials and procedures were revised based on the findings of the pilot test and finalized for the experiment.

Criterion measures

The primary criterion measure of performance for each student was the score on the final lesson plans. The draft lesson plans were also scored by the researcher after the study to provide comparative data between the draft and final versions. Both versions were scored blind—that is, without knowledge of each subjects' treatment group.

Student scores on a 15-item posttest served as a measure of knowledge. The criterion measure for student attitudes was the student attitude survey administered after the posttest. Teacher interviews were conducted using a teacher-interview protocol which contained questions about the teacher's perceptions of the evaluation processes.

Final lesson plan scores

The 15-item lesson-plan rubric distributed to all students in week 2 served as a guide to them in development of their lesson plans. The rubric was based on the presence or absence of the appropriate content and the quality of the content of a good lesson plan that integrates technology into a lesson. The rubric was reviewed for content validity by three subject-matter experts during the pilot study, then revised into final form based on their feedback.

The rubric involved 15 criteria and was organized into four sections. (1) standards and objectives, (2) technology and materials, (3) lesson procedures, and (4) assessment. Each criterion in the rubric was rated on a scale from 0 to 2.

Thus the maximum possible score on a lesson plan was 30. A sample item from the lesson plan evaluation rubric is provided below:

Lesson procedures are aligned with technology standards.

0. Lesson procedures are not aligned with technology standards.
1. Lesson procedures are partly aligned or the teacher, instead of the student, performs the procedures covered in the standards.
2. Lesson procedures are aligned with technology standards and student performs the standards.

Students in each evaluation group were given the time from week 4 to week 5 to finalize and submit their lesson plans after receiving the week 4 feedback on their draft lesson plans. The final lesson plans were scored by the primary researcher and one independent rater trained by the researcher. Both raters were unaware of the experimental condition under which each plan was developed. After training on a sample set of plans, the Pearson correlation coefficient for interrater reliability between the ratings of the two raters on the lesson plans was .93.

The instructors scored the final lesson plans for grading purposes. The Pearson correlation coefficient for interrater reliability between the ratings of the researcher and the teachers was .83. The reliability of the rubric, using Cronbach's alpha, was .84.

Posttest

The posttest served as a measure of student knowledge. The content validity of this multiple-choice test was addressed by directly aligning one test item each with an item on 15-item lesson-plan rubric. Internal reliability of the posttest, using Cronbach's alpha, was .53, a low reliability coefficient due at least in part to the low number of items on the test. A sample multiple-choice item from the posttest is provided below:

Which state standard best aligns with the objective below?
Objective: The students will use input and output devices to operate the computer successfully.

A) Understands the operations and function of technology systems and is proficient in use of them.
B) Uses technology-drawing tools for communicating and illustrating.
C) Demonstrates respect for other students while using technology.
D) Creates criteria to compare and contrast technology systems, resources, and services.

Student attitude survey

The student attitude survey consisting of seven Likert-type items and four open-ended questions served as the criterion measure for assessing student attitudes and continuing motivation toward the evaluation conditions. The

internal reliability of the student attitude survey, using Cronbach's alpha, was .82. The Likert-type items were statements with four response choices: *strongly agree*, scored as 4, *agree* (3), *disagree* (2), and *strongly disagree* (1). These items consisted primarily of statements addressing student attitudes toward their form of evaluation and its influence on their lesson plans. The open-ended questions dealt with such topics as who the students would like to have as their evaluator and the perceived strengths and problems related to their assigned evaluation condition.

Teacher interview protocol

Individual interviews with the teachers were conducted by the primary researcher using a structured interview protocol. The interview protocol was developed to capture teacher attitudes and observations related to the three evaluation conditions. The interviews were conducted with the teachers after the experimental part of the study was completed. The interview protocol included questions covering all of the three evaluation conditions. Initially teachers were asked to identify the two evaluation types they conducted. The interviewer asked each teacher only the questions related to those two evaluation types. The interview protocol contained standard questions for each evaluation condition about the strengths of the condition, the problems encountered with it, its effect on the final lesson plans, and the teachers' thoughts about using it for formative-evaluation purposes. Each interview session lasted approximately 20 min. The researcher recorded the key ideas from each interviewee's responses in writing during the interview.

Data analysis

The primary data analysis for student performance on the lesson plans was a 3 treatment: (teacher-evaluation, peer-evaluation, self-evaluation) $\times$ 2 lesson plan version (draft lesson plans and final lesson plans) repeated-measures analysis of variance (ANOVA). Treatment was a between-subjects variable and lesson-plan version a within-subjects variable. Based on the results of this ANOVA, follow-up analyses were conducted to analyze performance by treatment groups across and within lesson-plan versions.

Student achievement on the posttest was analyzed by using a one-way ANOVA across the three evaluation types, followed by Scheffe tests to identify significant between-group differences. En route data on the number of comments on the draft lesson plans by each of the three types of evaluators were also analyzed using ANOVA. Student attitude data were analyzed by multivariate analysis of variance (MANOVA), followed by univariate analysis of the mean scores on each survey item.

The qualitative data from the teacher interviews and student responses to open-ended questions were summarized using procedures described in Miles and Huberman's (1984) qualitative data-analysis model. These procedures, as applied in the present study, included identifying patterns or themes in the

responses, summarizing and tabulating the responses under each pattern, comparing the responses across the three evaluation conditions, and drawing conclusions from the patterns and comparisons.

Results

Results are reported in this section for lesson-plan ratings, posttest scores, amount and type of feedback under each evaluation condition, student attitudes, and teacher interviews.

Lesson plan ratings

The mean scores and standard deviations for both the draft lesson plans and final lesson plans scored by the researcher are shown in Table 1. The table reveals that the mean scores for the draft lesson plans were 16.92 out of 30 possible (56%) for the teacher-evaluation group, 17.55 (59%) for the self-evaluation group, and 17.72 (59%) for the peer-evaluation group. Final lesson plan scores were 24.78 out of 30 possible (83%) for the teacher-evaluation group, 20.07 (70%) for the self-evaluation group, and 20.94 (70%) for the peer-evaluation group.

The 3×2 repeated-measures ANOVA for the researcher-scored lesson plans yielded a significant main effect for lesson-plan version, F (1,98) = 77.14, p = <.01. This significant difference is reflected in the fact that the scores on the draft plans by the three groups were in the 16.9–17.7 range, while the scores on the final lesson plans ranged from 20.1 to 24.8. The ANOVA did not yield a significant main effect for evaluation type, F (2,98) = 2.47, p = .09, which was based on the combined scores across the two lesson-plan versions for each evaluation group.

The 3×2 repeated-measures ANOVA also yielded a significant evaluation type by lesson-plan version interaction, F (2, 98) = 10.82, $p < .01$. This interaction reflects the fact that the draft lesson-plan scores of the three evaluation types were quite similar with the teacher-evaluation score being slightly the lowest, whereas the final lesson-plan score for the teacher-evaluation group (24.78) was several points higher than those of the self-evaluation (20.07) and

Table 1 Draft and final lesson plan scores by evaluation type

Measure	Evaluation type			
	Teacher evaluation	Self-evaluation	Peer evaluation	Overall
Draft lesson plans				
M	16.92	17.55	17.72	17.39
SD	(5.15)	(4.04)	(3.30)	(4.22)
Final lesson plans				
M	24.78	20.07	20.94	22.06
SD	(3.78)	(5.26)	(5.96)	(5.42)

Note: Maximum possible points on each lesson plan was 30

N = 36 for teacher evaluation, N = 29 for self-evaluation, N = 36 for peer evaluation

peer-evaluation (20.94) groups. The evaluation type by lesson-plan version interaction is diagrammed in Fig. 1.

Paired-sample t tests were conducted to identify significant differences between pairs of groups in the significant evaluation type by lesson-plan version interaction. All three evaluation groups scored significantly higher on their final lessons than they had scored on their own draft plans: teacher, t (35) = 9.77, $p < .01$; self, t (28) = 2.92, $p < .01$, and peer, t (35) = 3.32, $p < .01$. Further, there were significant differences in the final lesson plan scores among the groups F (2, 98) = 8.29, $p < .01$. The follow-up Scheffe test revealed that 24.78 score for the teacher-evaluation group on the final lesson plans was significantly higher than both the 20.07 score for the self-evaluation group and the 20.94 score for the peer-evaluation group. None of the scores on the draft lesson plans differed significantly from one to another.

Posttest scores

The mean posttest scores were 9.75 out of 15 (65%) for the teacher-evaluation group, 10.34 (69%) for the self-evaluation group, and 8.11 (55%) for the peer-evaluation group. The ANOVA for posttest scores yielded a significant overall difference, F (2, 98) = 7.81, $p < .01$. The follow-up Scheffe test revealed that subjects in both the teacher-evaluation group and the self-evaluation group scored significantly higher at the $p < .01$ level than those in the peer-evaluation group. The difference in posttest scores between the teacher-evaluation group and the self-evaluation group was not significant.

Draft lesson plan scores by evaluators

As part of the evaluation of the draft lesson plans, the three types of evaluators used the 15-item standard rubric and rated each of the 15 criteria from 0–2, resulting in a maximum possible score of 30 on each lesson plan. The overall mean scores on the draft plans, as scored by the three types of eval-

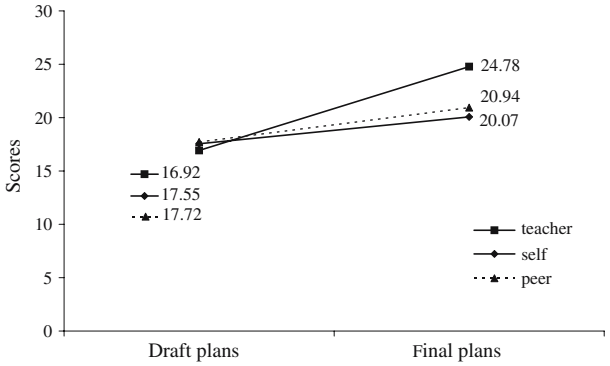

Fig. 1 Evaluation type by lesson–plan version interaction

uators, were 16.44 under teacher-evaluation, 22.90 under self-evaluation and 20.94 under peer-evaluation. Univariate analysis of variance on these draft lesson plans scored either by the teacher, oneself (the lesson-plan writer) or a peer revealed a significant overall difference among the three groups, $F(2, 98) = 9.60, p < .01$. Post hoc Scheffe tests indicated that the mean ratings assigned by both peer and self-evaluators were significantly more positive than the ratings assigned by the teacher at the $p < .01$ level. The difference in the draft lesson plan scores between the peer-evaluation and self-evaluation group was not statistically significant.

Amount and type of feedback

All three types of evaluators were asked to provide feedback-type comments for the lesson plan writer on at least four of the 15 criteria included in the rubric. Peer evaluators provided comments on an average of 6.2 items, teacher evaluators on an average of 5.2 items, and self-evaluators on an average of 4.7 items. Analysis of variance on the number of comments revealed a significant overall difference across the type of evaluators $F(2, 98) = 3.90, p < .05$. A post hoc Scheffe test for pairwise comparisons indicated that the number of feedback-type comments provided by the peer-evaluation group was significantly higher than the number provided by the self-evaluation group ($p < .05$). The number of teacher comments did not differ significantly from those of the other two groups.

The type of feedback comments was analyzed across treatment conditions using Miles and Huberman's (1984) qualitative data analysis model. Teacher comments on 32 of the draft lesson plans contained corrective feedback, usually in the form of pointing out poorly designed parts of the lesson plan and indicating how to improve these parts. Examples of such teacher feedback included ''Objectives are not observable. Think about how you can observe that they understand the concept.'' and '' Technology standard is not aligned. Look for one that is for using technology as a research tool.'' Twenty-four of the self-evaluation rubrics contained comments that were either mostly positive about the quality of the work or contained a reminder to fix a certain part, such as ''Good,'' or ''Add technology standards,'' or ''Add assessment rubric.'' Peer-evaluator comments were more informative in nature than comments in self-evaluations, telling the peer what was missing but typically not providing information about how to fix it. Examples of such comments included ''Assessment is not aligned with objectives,'' or ''Rewrite objective 1 and 2,'' or ''Good objectives, but from the objectives I have no idea what students will learn.'' Peers also provided comments citing the good parts of the work, such as ''Good lesson plan,'' ''Unique idea that will make the kids interested,'' or ''Great idea.''

Teachers reported that on average they spent 10 min on scoring and providing feedback to each lesson plan. Peer evaluators reported spending an

average of 13 min on evaluating the assigned lesson plan, and the time self-evaluators reported spending an average of 8 min.

Student attitudes

The items in the attitude survey, which was administered after students submitted their final lesson plans, were scored on a four-point scale from 4 for strongly agree to 1 for strongly disagree. It can be seen from Table 2 that the overall mean scores across the seven attitude items were 3.30 (out of a possible 4) for teacher-evaluation, 3.28 for self-evaluation and 3.04 for peer-evaluation. These scores show that student attitudes toward all three formative-evaluation strategies used in the study were favorable. Students in all three evaluation groups agreed or strongly agreed with the statements ''I made some changes in my lesson plan because of the evaluation'' and ''The feedback from evaluation helped me to prepare a better lesson plan.''

A 3 (treatment) × 7 (survey items) MANOVA performed on the attitude items revealed that the overall means were significantly different across the three treatment groups, Wilks' $\Lambda = .782$, $F\ (14, 186) = 1.74$, $p < .01$. Pair-wise comparisons indicated that the overall mean score for teacher-evaluation was significantly higher than the mean for peer-evaluation at the $p < .05$ level. The differences between teacher-evaluation and self-evaluation and between self-evaluation and peer-evaluation were not statistically significant.

Follow-up univariate analyses of variance conducted on each of the seven items revealed significant attitude differences beyond the .05 level on two of the seven items. Students under teacher-evaluation rated the items ''I made some changes in my lesson plan because of the evaluation'' and ''I liked the type of evaluation I got'' significantly higher than those under peer-evaluation.

Table 2 Student attitudes by evaluation type

	Treatment				
	Teacher	Self	Peer	*F*	*p*
1. I liked doing the evaluation of the lesson plan.	3.25	3.28	3.06	1.71	ns
2. I received good directions for conducting the evaluation.	3.22	3.28	2.97	2.13	ns
3. I would like to be evaluated like this again.	3.22	3.00	3.00	1.39	ns
4. It was easy for me to use the evaluation rubric.	3.13	3.34	3.00	2.19	ns
5. I liked the type of evaluation I got.	3.22	3.10	2.81	3.07	< .05*
6. I made some changes in my lesson plan because of the evaluation.	3.64	3.52	3.22	5.06	< .01*
7. The feedback from evaluation helped me to prepare a better lesson plan.	3.42	3.41	3.22	1.16	ns
Overall means	3.30	3.28	3.04	1.74	<.01*

* Teacher significantly higher than peer on these three comparisons. No other significant differences

One open-ended question asked if students would like to be evaluated by someone other than their assigned evaluator. Twenty-one students in the self-evaluation group (72%) and 16 in the peer-evaluation group (44%) reported that they would like to be evaluated by someone else, compared to only four in the teacher-evaluation group (11%). The most common preference for both the self-evaluation subjects (34%) and peer-evaluation subjects (14%) was to have the teacher as the evaluator. The next most common preference for the self-evaluation subjects was to have a peer as the evaluator (24%), whereas the next common preference of peer-evaluation subjects was either to have a different peer as the evaluator (11%) or to have themselves as the evaluator (11%). Five self-evaluators and six peer-evaluators also mentioned that they wanted to provide more positive feedback to themselves under self-evaluation or to their peers under peer-evaluation. This indicated a desire to provide feedback that made the lesson-plan writer feel positive about his/her work, even if their overall ratings were not particularly positive.

In their responses to open-ended questions regarding the problems encountered, peer and self-evaluators commented from two different perspectives: being evaluated themselves and being the evaluator. Ten peer-evaluation subjects and 12 self-evaluation subjects responded by stating the problems they encountered as the evaluator, and 43 responded from the perspective of being the person who was evaluated.

Students were also asked about the strengths and the problems related to their assigned evaluation condition. The most common strength reported was ''being able to fix the parts that were missing'' for the teacher-evaluation group (5 students), ''double check of the work'' for the self-evaluation group (6 students), and ''feedback from a peer'' for the peer-evaluation group (7 students). The most commonly reported problem was ''teacher feedback not very detailed'' for the teacher-evaluation group (3 students), ''lack of outside feedback'' for the self-evaluation group (7 students), and ''credibility of the evaluator'' for the peer-evaluation group (3 students). Several students also mentioned the lack of training on use of the evaluation rubric as a weakness of both the self and peer processes.

Teacher interviews

The individual interviews with the teachers revealed that they liked the use of peer and self-evaluation in the formative-evaluation process. Each of the three teachers mentioned that the self and peer procedures eliminated the major errors in the lesson plan before it was submitted to the teacher. Each teacher also stated that both self- and peer-evaluation enabled students to gain a better understanding of the scoring rubric.

The teachers were asked in their interviews to give the strengths and describe the problems associated with the three evaluation procedures. They cited their own ability to evaluate the lesson plans and provide corrective feedback as the major strength for the teacher-evaluation process, enabling students to critically review their own work and enhance their perception of

self-worth as evaluators as strengths of self-evaluation, and engagement with the scoring rubric and opportunity to see the quality of others' work as strengths of peer-evaluation. One teacher stated that self-evaluation is a good personal skill for the future because it shows students that, if they take the time, they can do better. As the main problems, for teacher-evaluation they cited the amount of time required for the teacher to do the evaluation, and for peer and self-evaluation they cited lack of experience with the evaluation process. Teachers also mentioned that self- and peer-evaluation were new experiences for students, so the students needed practice on use of the evaluation rubric and on performing the evaluation and feedback tasks.

Discussion

This study investigated the effects of teacher-evaluation, self-evaluation and peer-evaluation as formative-evaluation strategies. Students in all three evaluation groups improved their performance significantly from draft lesson plans to final lesson plans. A significant evaluation type by lesson-plan version interaction also revealed that students in the teacher-evaluation group produced final lesson plans that were rated significantly higher than the final plans of students in the self-evaluation and peer-evaluation groups. On the posttest measure, students in the teacher-evaluation and self-evaluation groups scored significantly higher than students in the peer-evaluation group. Also the attitudes of students in the teacher-evaluation and self-evaluation groups were generally more positive toward their treatments than those of students in the peer-evaluation group. Students receiving teacher-evaluation reported that they made changes in their lesson plans more frequently than did those in the peer-evaluation group.

While the findings from the teacher-evaluation group were generally more positive, the finding that subjects in the self-evaluation and peer-evaluation groups made significant improvements from the draft versions to the final version was also encouraging. These treatments required students, who could be considered to be generally inexperienced as evaluators, to use the evaluation rubric to make judgments about the quality of their own or a peer's lesson plan. That they were at least somewhat successful at this task is reflected in the significant improvement in performance under these two conditions.

The results favoring teacher-evaluation over self- and peer-evaluation for writing lesson plans were most likely due to the greater knowledge of the teachers about writing the plans and their experience in evaluating them. The greater knowledge and experience enabled the teachers to provide better feedback on the student lesson plans. Also, students may have considered the teacher feedback to be more important and may have responded more conscientiously to it because the teacher would be determining their final grade. The students in the teacher-evaluation group reported that they made changes in their lesson plans to a significantly greater extent than those in the peer

group and to a slightly greater extent than those in the self-evaluation group. The cycle of better feedback followed by more revisions appeared to have the effect of producing higher-quality lesson plans.

Several sources of data from the study support the idea that teacher-evaluators provided better feedback on the draft lesson plans and that this feedback was instrumental in helping students improve their plans. The significantly lower scores assigned by teacher evaluators than by self-evaluators and peer evaluators on the draft lesson plans suggest that the teachers were able to detect more weaknesses in the draft lesson plans than the other two groups. Analysis of comments on the draft lesson plans revealed that the teacher evaluators provided corrective feedback more often than the other two groups. In addition, in their interviews teachers cited their ability to evaluate the draft lesson plans and provide constructive feedback as the major strength of the teacher-evaluation. Students in the teacher-evaluation group also reported that ''being able to fix the parts that are missing'' as the most common strength of teacher-evaluation. Many more students in both non-teacher-evaluator groups than in the teacher-evaluator group expressed a preference for being evaluated by someone other than their assigned evaluator, and the most common preference as an evaluator for both these non-teacher groups was the teacher.

Not only did the students recognize the teachers' ability to provide better corrective feedback on their draft lesson plans, many of them also indicated feelings of inadequacy for conducting the assigned evaluation task. Their concern was apparent in the reports by the self-evaluation group that ''lack of outside feedback'' was the major weakness of their treatment and by the peer-evaluation subjects that ''credibility of the peer evaluator'' was a weakness of their treatment. Several peer evaluators reported in their responses to open-ended questions that they tried hard to provide good feedback, but they lacked the experience to do it. This lack of experience, especially on the draft lesson plans in cases where peer feedback may have been perceived by the recipient as not being particularly helpful, was a likely factor contributing to the credibility issue and to the less positive attitudes toward peer-evaluation. As Brindley and Scoffield (1998) noted, students may view peer-evaluation less positively because they do not consider their peers knowledgeable or experienced enough to provide them with helpful feedback.

Two factors appear to have contributed to the significantly higher scores assigned on the draft lesson plans by the student evaluators than by the teachers. One is the students' lesser knowledge about writing and evaluating lesson plans, and the other is their desire to provide more positive feedback to themselves and to their peers. Students received instruction on the criteria in the evaluation rubric prior to writing their draft lesson plans, but they did not receive training in use of the rubric for evaluation purposes. In retrospect, it seems likely that such training would help student evaluators improve their evaluations and feedback. For example, data from this study suggest that students may benefit from preliminary practice on using the rubric to evaluate one or more sample lesson plans containing common errors, as well as from

instruction that addresses the use of corrective feedback in evaluation and the desire by students to provide overly positive feedback to themselves and peers.

The fact that all three-evaluation groups provided an average of more than four written feedback-type comments on the rubrics undoubtedly was influenced by the statement at the beginning of the rubric telling them to provide written feedback on at least four criteria. It is clear that this statement generally had the desired effect because fewer than 10% of the evaluators failed to provide at least four comments. Our experience with formative-evaluation of student work indicates that it is advisable for researchers in this area to provide the evaluators with guidance regarding the minimum number of feedback comments to make.

The peer-evaluators provided more feedback comments on the draft lesson plans than either of the other two groups and significantly more than the self-evaluators. In their interviews, the teachers emphasized that peer evaluators felt accountable and tried hard to provide good feedback. It is possible that self-evaluators provided fewer written comments to themselves because they knew that they would see their own lesson plans again and may have felt that it was not necessary to note all the possible weaknesses or corrections on the draft plan. The teacher and peer evaluators, on the other hand, had only their one opportunity to provide comments and information to the lesson-plan writers.

Students in the self-evaluation group, as well as those in the teacher-evaluation group, scored significantly higher on the 15-item knowledge posttest than those in the peer-evaluation group, even though the performance of self-evaluation and peer-evaluation groups did not differ significantly on the lesson plans. The reason for this difference on the posttest between the two student-evaluation groups is not clear. The posttest was directly aligned with the lesson plan evaluation rubric, with one item in the 15-item posttest being related to each item on the 15-criteria rubric. Both the self-evaluation and peer-evaluation groups were presented with the rubric prior to initial development of their draft lesson plans and used it during their assigned evaluation task and presumably during revision of their own plans. Thus, the most logical explanation would seem to be that the two student groups would not differ in their knowledge related to the rubric. Perhaps the significantly higher score favoring the self-evaluation group over the peer evaluators may have been related to self-evaluators' comments about the value of being required to double check their own work and/or to the credibility issue with peer evaluators.

Overall, teacher-evaluation was the most effective of the three evaluation procedures used in this study for helping students improve their performance in developing good written lesson plans. However, a major concern expressed by the teachers about the teacher-evaluation process was the amount of time required for them to do good formative evaluation of the students' work and to provide useful feedback to the students. The teachers reported spending an average of 10 min on each lesson plan, or an average of nearly three hours per

section, to evaluate the plans and provide formative feedback. They were willing to do this because of an established relationship with the lead researcher and with her program area at the university, but many teachers may be unwilling to spend so much time on en route evaluation of student performance, especially if they also must subsequently review the final version of each student's product for grading purposes.

The results for the self- and peer-evaluation conditions were quite consistent with several of the potential benefits cited by proponents of these procedures and reported in the introductory section of this manuscript. These potential benefits included improved student performance, a more active student role in their own learning, a greater awareness of the evaluation process and scoring criteria, and an increased understanding of the instructional content (Brindley & Scchofield, 1998; Smith et al., 2002; Davies, 2000, 2002; Fallows & Chandramohan, 2001). Although participants achieved the greatest improvement in their lesson plans under formative evaluation by their teacher, subjects played a more active role in their own learning and also made significant improvements in their plans under both student-evaluation conditions. In addition, all three teachers reported that both self- and peer-evaluation enabled students to gain a better understanding of the evaluation rubric and to eliminate the major errors in their lesson plans. Also, students under both student-evaluation conditions agreed quite strongly on the attitude survey that they made changes in their lesson plan because of the evaluation and that the evaluation helped them to prepare a better lesson plan.

Although teacher-evaluation was the most effective procedure overall, the results were also quite promising for self-evaluation and peer-evaluation, especially if improvements can be made in these procedures. Having students evaluate their own work or the work of a peer based on a set of criteria resulted in better performance on the student work. Direct training of students on the use of the rubric for the purpose of formative evaluation, which was not explicitly included in the present study, is a potential improvement that could be incorporated into future research. As noted earlier, effective training of this type could be directed toward such factors as helping student evaluators to detect more errors in student work, provide more constructive feedback, and avoid feedback that is excessively positive.

Additional observations from the present study may be used to guide the design of formative-evaluation procedures in the classroom. Because the teacher and student techniques have different strengths and weaknesses, it may be desirable to use them in combination for formative purposes in order to capitalize on their strengths and minimize their weaknesses. Teachers are typically more knowledgeable about lesson plans and other student learning tasks. Thus, they often can provide students with more specific improvement suggestions. At the same time, teachers in this study indicated that they could not afford to routinely provide such formative feedback due to the significant time commitment. To save time, teachers may want to consider reviewing only a selected set of student assignments, or relying more heavily on their previous

experience with the learning task, to provide whole-class feedback regarding the most typical errors that students make. Students could then engage in self and/or peer-evaluation and double-check their work in response to the teacher feedback.

Further research could help to inform several areas related to classroom formative-evaluation practices. Research on procedures for training students on the evaluation criteria and on effective feedback for a task could improve their performance as self and peer evaluators. Studies aimed at identifying the best combination and sequence for using teacher and student evaluation together would also be helpful. In addition, the question of whether engagement in self- and peer-evaluation improves student self-evaluation ability is an important one. Future research that addresses issues such as these should help us identify effective formative-evaluation practices for improving student work on performance-based tasks.

References

Andrade, H. G., & Boulay, B. A. (2003). Role of rubric-referenced self-assessment in learning to write. *Journal of Educational Research, 97*(1), 21–32.

Ballantyne, R., Hughes, K., & Mylonas, A. (2002). Developing procedures for implementing peer assessment in large classes. *Assessment and Evaluation in Higher Education, 27*(5), 427–441.

Black, P., & William, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy and 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.

Brindley, C., & Scoffield, S. (1998). Peer assessment in undergraduate programs. *Teaching in Higher Education, 3*(1), 79–89.

Cheng, W., & Warren, M. (1997). Having second thoughts: Student perceptions before and after a peer assessment exercise. *Studies in Higher Education, 22*(2), 233–239.

Chi, M. T. H., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science, 13*, 145–182.

Davies, P. (2000). Computerized peer assessment. *Innovations in Education and Training International, 37*(4), 347–355.

Davies, P. (2002). Using students reflective self-assessment for awarding degree classifications. *Innovations in Education and Teaching International, 39*(4):307–319.

Doering, A., Hughes, J., & Huffman, D. (2003). Preservice teachers: Are we thinking with technology? *Journal of Research on Technology in Education, 35*(3), 342–361.

Elawar, C. 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*(2), 162–173.

Falchikov, N. (1986). Product comparisons and process benefits of collaborative peer group and self assessments. *Assessment and Evaluation in Higher Education, 11*(2), 146–166.

Falchikov, N., & Boud, D. (1989). Student self-assessment in higher education: A meta-analysis. *Review of Educational Research, 59*(4), 395–430.

Fallows, S., & Chandramohan, B. (2001). Multiple approaches to assessment: Reflections on use of tutor, peer and self-assessment. *Teaching in Higher Education, 6*(2), 229–245.

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

Fontana, D., & Fernandes, 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.
Kitsantas, A., & Baylor, A. (2001). The impact of the instructional planning self-reflective tool on preservice teacher performance, disposition, and self-efficacy beliefs regarding systematic instructional planning. *Educational Technology Research and Development, 49*(4), 97–106.
Liu, Z. F. E., Lin, S. S. J., & Yuan, S. M. (2002). Alternatives to instructor assessment: A case study of comparing self and peer assessment under a networked innovative assessment procedure. *International Journal of Media, 29*(4), 395–403.
Miles, M. B., & Huberman, A. M. (1984). *Qualitative data analysis: A source book of new methods*. Beverly Hills, CA: Sage.
Miller, L., & Ng, R. (1994). Peer assessment of oral proficiency. *Working papers of the Department of English City Polytechnic of Hong Kong, 6*:41–56.
Olina, Z., & Sullivan, H. (2002). Effects of classroom evaluation strategies on student achievement and attitudes. *Educational Technology Research and Development, 50*(3), 61–75.
Olina, Z., & Sullivan, H. (2004). Student self-evaluation, teacher evaluation and learner performance. *Educational Technology Research and Development, 52*(3), 5–22.
Reiser, R. A. (1994). Examining the planning practices of teachers: Reflections on three years of research. *Educational Technology, 34*(3), 11–16.
Roberts, S. K., & Hsu. Y. S. (2000). The tools of teacher education: Preservice teachers use of technology to create instructional materials. *Journal of Technology and Teacher Education, 8*(2), 133–152.
Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18*, 119–144.
Scriven, M. (1967). The methodology of evaluation. In R. Tyler, R. Gagne, & M. Scriven (Eds.), *Perspectives of curriculum evaluation: Vol. 1. AERA Monograph series on curriculum evaluation* (pp. 39–83). Chicago: Rand McNally.
Smith, H., Cooper, A., & Lancaster, L. (2002). Improving the quality of undergraduate peer assessment: A case for student and staff development. *Innovations in Education and Teaching International, 39*(1).
Stiggins, R. J. (2001). *Student-involved classroom assessment*. Upper Saddle River, NJ: Merrill Prentice Hall.
Topping, K. J., Smith, E. F., Swanson, I., & Elliot, A. (2000). Formative peer assessment of academic writing between postgraduate students. *Assessment & Evaluation in Higher Education, 25*(2), 149–167.
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.
Zhang, S., (1995). Re-examining the affective advantages of peer feedback in the ESL writing class. *Journal of Second Language Writing, 4*(3), 209–222.

Gamze Ozogul is a doctoral candidate in the Educational Technology Program at Arizona State University.

Zane Olina is an Assistant Professor of Instructional Systems at Florida State University.

Howard Sullivan is Professor in the Educational Technology Program at Arizona State University.