In-Person Grading: An Evaluative Experiment

J. Philip East
Computer Science Department
University of Northern Iowa
Cedar Falls, IA 50614-0507
319-273-2939
east@cs.uni.edu

J. Ben Schafer
Computer Science Department
University of Northern Iowa
Cedar Falls, IA 50614-0507
319-273-2187
schafer@cs.uni.edu

ABSTRACT
In this paper, we discuss in-person or face-to-face grading: what it is, a rationale for its use, our use of it, and an experiment we conducted to evaluate its use. While no statistically significant differences in instructional outcome effects were found, several interesting affective results were seen. Additionally, a number of research methodological suggestions arose from the study.

Categories and Subject Descriptors
K.3.2 [**Computer and Information Science Education**]: Computer science education

Keywords
CS Ed research, classroom management, face-to-face grading, in-person grading, pedagogy, providing programming feedback

1. INTRODUCTION

1.1 Background
Conscientious computer science educators reflect on their practice and seek to improve student learning. Teaching consists of many activities including course planning; lesson planning, delivery, and assessment; assignment planning, assistance, and grading; exam preparation, administration, and grading; and management of various resources. Thus, there are many possibilities for improving instruction and learning. When students make errors or have difficulty, providing feedback is key to enhancing learning. In computer science education, a variety of activities for generating data for such feedback and approaches to communicating the feedback to students have been discussed in the literature ([7], [8]). However, we have found no prior research that actually examined the instructional effectiveness of alternatives for providing feedback to students on their programming assignments.

Ruehr and Orr [8] identify five approaches to providing feedback on students' programming assignments—instructor notes on submitted programs, automated grading, peer review, public presentation, and interactive program demonstration. Cooper [2] and others [1, 5] refer to the latter as face-to-face grading and we have referred to it as in-person grading ([3], [4]). The technique uses personal and private meetings between instructor and student to discuss the student's work and the instructor's evaluation of it. Such meetings may last 20 to 30 minutes and, due to time constraints, may not occur for all assignments for all students.

The first author's introduction to in-person grading occurred via the SIGCSE discussion list in the summer of 2000. The discussion was quite positive and referenced Cooper who said "[face-to-face grading] is probably the single most important improvement that can be made in course management." ([2], p.130) Since then, our study of in-person grading has been a multi-stage research process and is culminating with the work reported here. The first step examined the practice's face-validity (see Kerlinger's discussion of construct validity [6] for general information and [3] and [4] for particular discussions). After concluding that the practice was reasonable theoretically (had face-validity) and was amenable to personal instructional style, the next step checked to see if in-person grading could be implemented by the particular instructor in specific courses [3]. That successful trial led to continued use of the practice, its introduction to the second author, and an eventual informal examination of how students viewed its use [4].

Combined positive results from theoretical, instructor, and student perspectives suggested more formal study was worthwhile. A literature search for prior research was essentially fruitless. The ACM Digital Library contained only two entries under "face-to-face grading" ([1] and [5]), both of which merely mentioned that the practice was not being used. A search for "in-person grading" yielded five entries, none of which actually addressed the subject practice. A search of the Web for the same terms produced about 115 possibilities, most of which were notices or instructions to students concerning someone's use of in-person grading. No actual research on the practice was found. The only actual research we found that related to grading practice was Olson's comparison of analytics and holistic grading of computer programs ([7]).

A more relaxed search of the Digital Library did produce a few results. A brief description of the practice can be found in [2] and more in-depth description and analysis of trade-offs is available in [8].

1.2 Rationale for the Study
Our early use of in-person grading had provided anecdotal evidence of the technique's usefulness. A theoretical analysis led to a similar conclusion with respect to both cognitive and affective outcomes. Several hypotheses were formulated.

Our personal teaching theories are constructivist—we believe student understanding is individually constructed from their experience (that their brains relate and integrate new knowledge and experience with prior knowledge and experience). Thus, they do not necessarily "learn" precisely what we "teach". To better determine student learning then we must attempt to determine the accuracy of their understanding. This cannot be done well by

SIGCSE'05, February 23-27, 2005, St. Louis, Missouri, USA.

merely examining program assignments and exam items. In-person grading sessions provide an opportunity to interact with students and question their understanding in order to identify and correct misunderstandings. If we are able to do that in a more frequent or more effective manner than with traditional feedback mechanisms, student learning should be improved.

In-person grading provides additional one-on-one interaction to any instructor's teaching and feedback practices. Several facets of instruction and learning should be enhanced by the increased personal attention afforded to students.

First, a more individual discussion of student work should have two positive effects even if instructors provide the same feedback in in-person grading that is given with other techniques. First, students are more likely to perceive and attend to comments from an instructor in the room with them than to written comments on a printout. Additionally, under in-person grading, students will have a much better opportunity to explore the meaning and implication of the feedback. In essence, the feedback should be more effective.

Second, one student activity believed to help learning is questioning. When students do not understand, instructors want them to ask questions. It is believed that the additional one-on-one and relatively informal time that students and instructor spend together under in-person grading will lead students to be more willing to ask questions. The increased questioning should occur both in the in-person grading sessions and during class time and should enhance student understanding and learning.

Finally, for many college instructors, the instructor-student relationship is relatively impersonal. The one-on-one time required for in-person grading is expected to personalize the instructor-student relationship and should lead to students having a better attitude toward the instructor and the class.

Specific hypotheses developed for the study were:

- student learning would be better under in-person grading than with other approaches to providing student feedback
- students being taught using in-person grading would ask more questions during class and, in general, participate in class at a higher rate than students under other feedback conditions
- students would prefer in-person grading over other feedback approaches

2. METHODOLOGY

This study was conducted in a relatively large comprehensive university (13,000 students) in the Midwest. The computer science department has eight faculty and approximately 200 undergraduate majors. In the fall of the 2003-2004 academic year three sections of CS I were taught by two instructors to 84 students.

Anticipated instructor effort (time) was planned and duties allocated in an attempt to divide the work appropriately according to the number of sections assigned each instructor and to control as many of the instructional variables as possible. The instructor assigned one section would do grading for all students except those participating in in-person grading and lead one of the weekly group feedback sessions. The instructor assigned two sections would “teach” (prepare and deliver lessons and prepare assignments), perform the in-person grading for about 10 students per assignment, and lead a second weekly group feedback session. Class time was on Monday, Wednesday, and Friday. Programming assignments were usually made weekly and were due on Friday. Students were required to submit a report for each assignment that discussed difficulties encountered, learning that occurred, questions on the assignment, and the time spent completing the assignment.

Three treatments (techniques for supplying feedback on programming assignments) were used—in-person grading, instructor-led discussion, and student-initiated discussion. Feedback was presented via personal in-person grading session or through group sessions held at a common evening time during the week. Students were randomly assigned to one of the three groups independent of lecture section.

Students in the student-initiated discussion group attended a Wednesday evening discussion session where the previous programming assignment was returned. Comments identifying problems, point deductions, and positive aspects were written on the program printouts. Students examined their assignments and were encouraged to ask questions regarding them. When no further questions were asked, the students were dismissed. Sessions generally lasted 30 minutes.

For instructor-led discussions, students also attended a Wednesday evening discussion and received their graded assignments. The program printouts had codes indicating problems seen and deductions made. The instructor made a presentation that explained the codes, discussed common problems, identified positive aspects of some student work, and provided alternative approaches to some parts of the assignment. (Codes were determined while grading programs for the student-initiated group and used in an attempt to reduce grading time.) Students were encouraged to ask questions at any time. When the presentation was over and no further questions were asked, the students were dismissed. Sessions lasted approximately one hour.

Students receiving in-person grading would sign up for a 20-minute session with the instructor to be held on Tuesdays. The instructor would examine each student's work and report prior to the appointment and prepare comments and questions for the student. Students were encouraged to ask questions. Only one-third of the in-person group received in-person grading on each assignment. In-person grading students were also randomly assigned to one of the other two groups and would attend the assigned discussion session when not undergoing in-person grading.

There were 10 assignments during the semester with the first being an introduction to the computer system, programming environment, and electronic submission system. Grading for it was not included in the experiment. Thus, each in-person grading student had three one-on-one sessions with the primary instructor and six group sessions.

A variety of data were collected during the semester. All student assignments and reports were kept, thus, the self-reports of time spent on each assignment were available for analysis. Student grades on assignments and exams were noted, as were course grades. Student attendance and participation in class and in the grading discussion sessions had been recorded. At the end of the semester a questionnaire was administered in an effort to determine student reaction and attitude about the experience. Elements of these data appropriate to our hypotheses were subjected to statistical tests of significance.

3. RESULTS

Based on feedback during our Human Subjects Review, students completing the course were asked for permission to include their course data and survey results in this study. The vast majority agreed to participate (only three students chose to withhold permission) with final numbers in each treatment appearing in Table 1.

Table 1 : Study Participants

Treatement	Number of Participants
Student Initiated (SID)	19
Instructor Led (ILD)	20
In Person Grading (IPG)	23
All Treatements	62

All results presented in this section are based on ANOVA tests including both the Tukey HSD and Bonferroni post hoc tests for statistical significance based on an $\alpha = 0.05$.

3.1 Course Results

From a pedagogical point of view we were interested in overall student performance in the course. Since students in the in-person treatment are required to interact with the professor in a one-on-one setting, we were interested to see if students in this treatment were more likely to participate in discussions either lecture or during group lab sections. Furthermore, we were interested in their performance in lab activities, the final exam, and the course overall. Table 2 contains the mean grades earned for students in each treatment for each of these five categories. While an initial examination of Table 2 may suggest that students in the student-initiated treatment performed best on each of these activities, the deviation in these grades is large enough to prevent any statistical significance in these results.

Table 2 : Mean course grade earned on various course activities. No statistical significance exists between mean grades earned for any of these activities.

[* SID=student-initaited discussion, ILD=instructor-led discussion, IPG=in-person grading]

Grade	SID*	ILD*	IPG*
Lab Participation	9.03	7.84	8.85
In-class Participation	5.82	5.39	5.2
Lab Scores	24.19	21.55	22.60
Final Exam	71.21	65.55	63.06
Overall Grade	73.22	66.48	68.50

3.2 Survey Results

While we were interested in the course-level outcomes resulting from using in-person grading, we were also quite interested in the affect on student attitudes toward the course. In order to consider such affects, we administered an end-of-course survey to all students concerning their comfort level, how closely they studied lab feedback, whether feedback was helpful, and their assessment of instructor efforts to improve feedback. Students were presented with a series of statements that they rated on a Likert scale from 1 (disagree strongly) to 5 (agree strongly) with 3 being a neutral score. The content of these statements and the mean score from members of each treatment are contained in Table 3.

Table 3 : Mean Likert score responses to end of course survey statements. Scores marked with an asterisk are different with statistical significance.

Num	Statement	SID	ILD	IPG
1	I feel more comfortable asking questions in this class than in many of my other classes.	3.68*	2.75*	3.48
2	I feel more comfortable going to office hours in this class than in many of my other classes.	3.94	3.35	3.87
3	I carefully examined feedback I received on assignments.	4.42	4.00	4.30
4	The feedback I received on programming assignments was useful.	3.32	3.25	4.04
5	Feedback on programming assignments affected the way I completed later programs.	4.37	4.10	4.57
6	The instructors were interested in my learning.	4.05	3.6	4.09
7	I approve of my instructors' attempts to improve instruction.	4.11	3.75	4.23
8	I would have preferred being in one of the other groups	3.20*	4.48*	1.72*
9	THIS TYPE OF feedback provided better feedback than typically encountered (in CS and non-CS classes).	3.37	3.05*	4.70*
10	THIS TYPE OF feedback improved my learning.	3.63	3.20*	4.22*
11	The instructors should continue using, and improving, THIS TYPE OF feedback.	3.74	2.90*	4.43*

As was the case with course performance, the differences in mean score were statistically insignificant for approximately half of the statements in the survey. However, several significant results were recorded.

Statement one asked students to indicate their level of agreement with the statement "I feel more comfortable asking questions in this class than in many of my other classes." It was our belief that students in the in-person treatment would have more interaction with the professor than students in the other treatments and would thus be more comfortable interacting with the professor. As can be seen in Table 3, the score for students in the in-person treatment is not statistically separable from that of students in the other treatments. However, students in the student-initiated

treatment indicate a significantly higher comfort level than students in the instructor-led treatment. In hindsight, this is perhaps explained by the fact that students in this treatment were somewhat forced to ask questions of the instructor. All discussion around performance on previous labs was based on the comments and questions from the students. They very quickly learned that they were expected to ask questions, and as such they would not be denigrated for such questions (likely a common fear among students). Furthermore, students in the instructor-led treatment were meeting with an instructor who was their instructor only for the lab discussion period. As such, they had limited interaction with the instructor over the course of the semester, and it is not surprising that they might never become sufficiently comfortable to freely ask questions.

Items nine, ten, and eleven all ask students to provide their evaluation on the feedback treatment they received. These consist of the statements, "THIS TYPE OF feedback provided better feedback than typically encountered (in CS and non-CS classes)," "THIS TYPE OF feedback improved my learning," and "The instructors should continue using, and improving, THIS TYPE OF feedback." Students receiving in-person feedback provided a much higher agreement ranking than their counterparts in the instructor-led sections. While it is likely that at least part of this difference can be explained by the "different instructors" factor mentioned previously, we suspect that this is largely due to the fact that students found in-person grading to be an effective and helpful method for providing feedback.

We make this previous conclusion based, in large part, on the results of scores received for statement eight. Reading, "I would have preferred being in one of the other groups" this statement was the only statement to produce mean scores that were statistically different for all pair-wise combinations. As can be seen in Table 3, students receiving in-person grading were very likely to disagree with this statement. In fact, with a mean score of 1.72, only three of the twenty-three student receiving in-person grading indicated they would prefer to receive some other form of feedback. This compares to average scores of 3.2 and 4.48 for students receiving student-initiated discussion and instructor-led discussion respectively. While the majority of students receiving in-person grading wanted to stay where they were, the majority of students receiving instructor-led discussion wished to change to a different group.

This data is confirmed by the results to a second part to statement eight. Students were given the option to indicate which treatment they would have preferred receiving. The results of this question are contained in Table 4. Again, it is clear that students receiving in-person grading were very happy where they were, students receiving instructor-led discussion were very unhappy where they were, and students receiving student-initiated discussion were somewhat equally split on their opinions.

Table 4 : Treatment students would have preferred receiving

Wished to be in…	Was in …		
	SID	ILD	IPG
SID	10	3	0
ILD	1	3	3
IPG	8	14	20

4. DISCUSSION

While we feel that in-person grading is a promising and effective feedback technique, we were disappointed that results did not more strongly support that belief. We have identified several factors we feel may have influenced the outcomes of this study.

First of all, we must acknowledge the overall weakness of the in-person grading "treatment". Students in this group met only three times over the course of the semester with their instructor. In a perfect setting in-person grading would occur for every lab assignment for every student. If, as we proposed, in-person grading is going to provide additional opportunities for students to ask questions, and instructors to evaluate student learning in a one-on-one setting, it must occur more frequently. In order for this to happen, however, student numbers must be kept reasonable and faculty must be dedicated to the concept of in-person grading. In-person grading for a class of 25 students requires a minimum of 8 hours of meeting times for each lab graded in this manner. We suspect that the majority of instructors are unable/unwilling to dedicate this amount of time to the process.

It is worth repeating, however, that even though student performance was not affected by this limited application of in-person grading, student attitudes were affected. As reported in the previous section, students in this treatment were far more satisfied with their feedback method than members of other groups, and were far less likely to indicate they wanted to switch to another group. Students appeared to recognize the usefulness of this one-on-one time even in its limited form. The following comment is similar to several comments received in the end of semester evaluations.

> "In person grading was an effective way to go through code, line by line and critique each aspect of my programming. It seemed to offer a far better evaluation than the [other groups]."

We also feel that, despite our best efforts to control variables, the results of this study were skewed due to differing instructor styles and the way we divided up the work between the two instructors. Members of the student-initiated discussion group interacted with a single instructor for both lecture and lab discussion portions of the course. Members of the instructor-led discussion group interacted with one instructor for the lecture portion of the course and a second instructor for the lab discussion portion of the course whose primary interaction with them was as grader or fault-finder. Based on student feedback on end of semester evaluations, this was clearly an issue that frustrated them. As one member of the ILD group wrote,

> "The only thing that I truly felt made a difference was having the [lab discussion] taught by a different professor than the classroom instruction. For me as a CS I student just learning to program, any differences in style or otherwise that the two instructors had, can be quite confusing. I think in a more experienced programming class that could be useful when developing your own style so that you could see two ways of doing it. But for CS I when basically fundamentals are the main concern for the course, it should have the same instructor for both lab and instruction."

While this may seem like it goes against what is standard practice at many universities—the use of professors for lecture instruction

and the use of TAs for lab instruction—we think there is a significant difference between this "standard" practice and what was experienced in our study. TAs are normally prepared for their role by the lecture professor and taught how to grade and what to recommend. When all else fails, most TAs defer to the lecture instructor. In this study however, we had two well-trained professors who differed in their approach to things and failed to communicate adequately the statements made in one context to maintain continuity with the other context. Additionally, early assignments were relatively easy and much grading focused on style elements not previously agreed upon by the instructors. This clearly was frustrating to students.

Needless to say, our future studies will strive to avoid the multiple-professor problem and to ensure sufficiently strong interventions. We are, however, pleased with this study. It has a solid and explicit theoretical foundation. It was well-designed with random assignment of students to treatments. We hope others can learn from our successes and failures.

5. CONCLUSIONS

Despite results that indicate there may be no effect on the final performance in the course for students participating in in-person grading, we feel strongly that this is a practice that we wish to continue to use, when appropriate, in our own instruction. The majority of students are very comfortable with this feedback technique and seem to feel it helps them learn how to be better programmers.

Furthermore, we feel this is a meaningful first step toward the design and implementation of courses that are nearly entirely laboratory based or "studio" courses. These would provide even more time for one-on-one interaction with students in an environment where students can immediately try the modifications that arise in discussions with the instructor and instructors can better evaluate the student comprehension of these discussions by evaluating the follow-up code written by students. Additionally, instructors would have a much better opportunity to get inside student heads to assess their actual understanding.

Regardless, we feel that in-person grading is a powerful evaluation and feedback tool that more instructors should consider incorporating into their own courses.

6. ACKNOWLEDGMENTS

The authors would like to thank the members of their department for assistance with this study. In particular, Dr. Eugene Wallingford who served as a third party administrator for participation agreements, course evaluations, and study surveys.

7. REFERENCES

[1] Astrachan, O., Smith, R, and Wilkes, J. Application-based modules using apprentice learning for CS 2. In *Proceedings of the Twenty-eighth SIGCSE Technical Symposium on Computer Science Education (SIGCSE '97)* (San Jose, California, February 27-March 1). ACM Press, New York, NY, 1997, 233-237.

[2] Cooper, D. Teaching Introductory Programming (with Oh! Pascal!). W.W. Norton & Company, New York, 1995.

[3] East, J.P. Experience with in-person grading. In Proceedings of the 34th Midwest Instruction and Computing Symposium (on CD-ROM). 2001. [Available on-line via http://MICSymposium.org]

[4] East, J.P. Experimenting with In-person Grading. Proceedings of the 37th Midwest Instruction and Computing Symposium (on CD-ROM). 2004 [Available on-line via http://MICSymposium.org]

[5] Kay, D.G. Large introductory computer science classes: Strategies for effective course management. In *Proceedings of the Twenty-ninth SIGCSE Technical Symposium on Computer Science Education (SIGCSE '98)* (Atlanta, Georgia, February 25-March 1). ACM Press, New York, NY, 1998, 131-134.

[6] Kerlinger, F.N. Foundations of Behaviorial Research (2nd Ed.). Hole, Rinehart and Winston, Inc., New York, NY, 1975.

[7] Olson, D.M. The reliability of analytic and holistic methods in rating students' computer programs. In *Proceedings of the Nineteenth SIGCSE Technical Symposium on Computer Science Education (SIGCSE '88)* (Atlanta, Georgia, February 25-26). ACM Press, New York, NY, 1988, 293-298.

[8] Ruehr, F. and Orr, G. Interactive program demonstration as a form of student program assessment. *Journal of Computing Sciences in Colleges, 18*, 2 (December, 2002), 65-78.