EDUCATIONAL ASSESSMENT, *11*(1), 1–31

The Impact of Self- and Peer-Grading on Student Learning

Philip M. Sadler
Science Education Department
Harvard-Smithsonian Center for Astrophysics

Eddie Good
Ephraim Curtis Middle School, Sudbury, MA

The 2001 U.S. Supreme Court Case of Falvo v. Owasso School System (Owasso Independent School District No I-011 v. Falvo) has focused national attention on the common classroom practice of peer-grading. In a unanimous decision the court reaffirmed the popular view that students grading each others' tests is valuable, saving teachers' time and augmenting student learning. Our study puts these presumed benefits to the test in 4 middle school science classrooms. We compared teacher-assigned grades to those awarded either by students to themselves or by their peers. By training students to grade with the help of a scoring rubric, a very high correlation was obtained between students and their teacher on test questions ($r = 0.91$ to 0.94). We found patterns of bias when students assigned grades. When grading others, students awarded lower grades to the best performing students than their teacher did. When grading themselves, lower performing students tended to inflate their own low scores. Performance on an unannounced, 2nd administration of the same test 1 week later measured the degree to which student-grading resulted in any increased understanding. Students who graded their peers' tests did not gain significantly more than a control group of students who did not correct any papers but simply took the same test again. Those students who corrected their own tests improved dramatically. Self-grading and peer-grading appear to be reasonable aids to saving teachers' time. Self-grading appears to result in increased student learning; peer-grading does not.

Teachers face a troubling conflict whenever they create tests and quizzes. On the one hand, they want assessments that measure the full variety of facts, skills, and

Correspondence should be addressed to Philip M. Sadler, Science Education Department, Harvard-Smithsonian Center for Astrophysics, Cambridge, MA 02138. E-mail: psadler@cfa.harvard.edu

concepts taught. On the other hand, their assessments need to accommodate limitations in resources and time. Items that measure sophisticated understandings require longer, more complex student responses. When given the opportunity, students write extensively, draw diagrams, create graphs, and provide examples from real life experiences in their tests and quizzes. In turn, responses that are more open-ended call for more time and effort on the teacher's part to read, correct, provide feedback, and grade fairly. For most teachers, increasing the sophistication of their assessment tools burdens them by leaving less time for other activities. This compromise plays out in the use of assessments that teachers feel are less than optimal but are easier to grade.

The critical resource is a teacher's time. At the middle and high school levels a single teacher can instruct more than 100 students while teaching four or five classes each day. In the United States, most middle and high school teachers must prepare for two different levels or types of course ("preps"), and some even more. For example, if a middle school science teacher with 100 students gives a unit test that may take 5 min per student to read and grade, this will add 8 hr of laborious correction time to his or her workload. Although standardized testing has become more prevalent in many states, this only accounts for 1% of classroom time (Phelps, 1997). Teacher-created testing can account for 10% or more of class time with students.

Student-grading, peer assessment, and self-assessment are terms that generally refer to "specific judgments of ratings made by pupils about their achievement, often in relation to teacher-designed categories" (Baird & Northfield, 1992, p. 21). Allowing students to grade tests themselves offers four potential advantages over teacher grading:

- *Logistical:* Because an entire classroom of students can be grading simultaneously, tests can be marked in a short amount of time. This saves teacher time (Boud, 1989). Grading can take place immediately following a quiz or during the next meeting of the class. This results in quicker feedback for students (McLeod, 2001). Peers can often spend more time and offer more detailed feedback than the teacher can provide (Weaver & Cotrell, 1986).
- *Pedagogical:* Judging the correctness of answers is an additional opportunity for students to deepen their understanding about a topic. Reading another's answers or simply spending time *pondering* another's view may be enough for students to change their ideas or further develop their skills (Bloom & Krathwohl, 1956; Boud, 1989).
- *Metacognitive*: Embedding grading as a part of a student's learning experience can have benefits that go beyond *learning* specific subject-matter content (Brown, 1987). Grading can help to demystify testing. Students become more aware of their own strengths, progress, and gaps (Alexander, Schallert, & Hare, 1991; Black & Atkin, 1996). Pupils develop a capacity to take initiative in evaluat-

ing their own work (Darling-Hammond, Ancess, & Faulk, 1995) and use higher order thinking skills to make judgments about others' work (Bloom, 1971; Zoller, 1993; Zoller, Tsaparlis, Fastow, & Lubezky, 1997). Self-evaluation and peer review are an important part of future, adult, professional practice, and test grading is a good way to develop these skills (Boud, 1989). With increasing awareness of the workings of tests, students can also formulate test items that can be used on later exams (Black & Harrison, 2001).

- *Affective*: Affective changes can make classrooms more productive, friendlier, and cooperative, and thus can build a greater sense of shared ownership for the learning process (Baird & Northfield, 1992; McLeod, 2001; Pfeifer, 1981; Weaver & Cotrell, 1986; Zoller, Ben-Chaim, & Kamm, 1997). The reason for tests is illuminated when students compare and judge the veracity of answers. Students develop a positive attitude toward tests as useful feedback rather than for "low grades as punishment for behavior unrelated to the attainment of instructional objectives" (Reed, 1996, p. 18; see also Sadler, 1989).

Viewed as a time-saving scheme, students' grading of their own or peers' work can only be considered a satisfactory substitute for teacher grading if the results of these grading practices are comparable to the teacher's. If student feedback or grades are very different from the teacher's judgment, the teacher is obligated to correct the papers herself. Ideally, student-assigned grades would be indistinguishable from grades assigned by the teacher. Although this may be easy to achieve when questions are of the form of multiple-choice or fill-in-the-blank, such agreement is more difficult in the case of more open-ended responses (Bloom & Krathwohl, 1956).

Because subjective judgment is often required for grading, particularly for more open-ended questions, students must learn how to correct tests (Boud, 1989; Neukom, 2000). Considered by many teachers as "guild knowledge," making judgments about students' understanding is a most arcane skill, acquired by apprenticeship and rarely revealed to the uninitiated (Sadler, 1989). A grading rubric or criteria sheet, which lays out different levels of response and equivalent point values, helps to specify a teacher's evaluative criteria and results in a greater agreement with a teacher's potential grade (Baird & Northfield, 1992; Boud, 1989; Weaver & Cotrell, 1986). This particular analytic approach of using a rubric depends on the teacher's effort in selection and codification of a subset of items from a large number of potential criteria that graders can consider separately but add up overall (Sadler, 1989).

Student-grading has several variations but only two major forms are addressed here. Students can correct either their own papers or those of others. Variations in peer-grading include the scoring of the work of students in one class by members of other classes and blind review with names undisclosed. Students can be rewarded for accuracy in grading of their own or another's work, with extra points earned if their judgments are similar to those of their teacher's.

BACKGROUND

The research literature concerning self- and peer-grading is substantive. Two meta-analyses provided a comprehensive list of quantitative self-assessment research (Falchikov & Boud, 1989) and peer assessment research (Falchikov & Goldfinch, 2000). An additional search for relevant items in EXXXX RXXXX IXXXX CXXXX and educational journals since 1969 uncovered only two additional quantitative studies (Doyle & Green, 1994; Zoller, Tsaparlis, et al., 1997). The self-assessment meta-analysis examined 57 studies and the peer assessment meta-analysis included 58 studies. All deal with college-age or older students. Also, no quantitative studies were found that attempted to measure the effect of student-grading on learning.

Boud (1989) remarked on the poor quality of quantitative research dealing with self-assessment, citing low technical quality and results and methods that vary from study to study. As noted in the literature, statistical methods used to compare the grades that students and teachers award include

- rank and Pearson correlation,
- agreement (also called proportion),
- Student's *t* test comparing difference in mean grade between teacher-assigned grades and student-assigned grades,
- effect size (difference in mean grades in units of standard deviation), and
- chi-square statistics (comparing grading categories).

The two meta-analyses attempted to remedy this situation by examining data from multiple studies. In general, individual studies are woefully lacking in details concerning their datasets, and it is difficult to reconstruct fully the methods used for analysis. Summary statistics generally substitute for graphical presentations of the data.

Methodologies

Although there are many possible approaches to designing studies and characterizing the difference between student- and teacher-assigned grades, little attempt is made in the literature reviewed to discuss the reasons for a particular design and the choice of measure, or to determine the error associated with a particular method. The two meta-analyses did discuss these issues.

The most obvious measure to use to compare the grades that two different judges use is "agreement," finding the fraction of student-assigned grades that coincide with the teacher's own (Burke, 1969). "Agreement" measures the exactness of a discrete match between teacher grading and student-grading. The percentage of exact matches (e.g., teacher and student assign a "B" to a student's paper) is the

statistic of interest. This measure is problematic because there are different definitions of agreement (Falchikov & Goldfinch, 2000). Researchers vary in using ±10 points difference, ± 1 standard deviation (on a 100-point scale), or an exact match of grade (on either a 5-category A, B, C, D, F scale or a 15-category scale using plus and minus grades in addition). Each of these measures would provide a different value for "agreement" using the same data, not a desirable situation.

The choice of analyzing student versus teacher grading using only such discrete agreement might arise from the presumed limited statistical background of many authors, some of whom are reporting on experiments in their history, writing, or psychology courses. Only the two meta-analyses mentioned the issue of generalizability of such a statistic and how the magnitude of such agreement would change with the coarseness of the grading scheme or the number of categories in these studies. Falchikov and Goldfinch (2000) pointed out that "draconian identical ratings" have lower agreement than within "ten marks [points] difference" (p. 293). Certainly for any individual teacher, this measure can act as a useful comparative indicator between identical grading schemes used in their own courses (i.e., whether one class better matches teacher grading than another). Falchikov and Goldfinch's meta-analysis recommended against agreement as a metric: "Future investigators would do well to avoid the use of proportions [i.e. agreement] as a common metric" (p. 293). They found 24 studies that used agreement as the statistic reflecting consistency between teacher and student assigned grades, whereas 56 studies used correlation.

It is easy to imagine how expanding grading categories from A, B, C, D, and F to a 100-point scale ranging from 1 (XXX) to 100 (XXX) could drop this agreement statistic to zero. Also, this measure does not take into account that there will be some agreement by chance, and even randomly assigned grades would have a few matches (Agresti & Findlay, 1997). Thus, the statistically adept might use Cohen's Kappa statistic to assess agreement between judges for nominal data and to account for chance agreement (Abedi, 1996; Fleiss & Cohen, 1973). Yet this still does not deal with the fact that with different numbers of categories in each scale, one gets different levels of agreement. Also, the use of discrete categories does not take into account the magnitude of individual differences between a student-awarded grade and a teacher-awarded grade. With strict agreement, a student grade of "B–" would not count as agreeing with a teacher grade of "B." However this "B–" is much closer to the teacher grade than a student grade of "F." Yet the agreement statistic does not statistically account for one being close and one being far from the teacher grade. A weighted Kappa is preferred for ordered data and takes into account the magnitude of the disagreement between raters. When ratings are ordinal, a ratio of variances by case and by judge can be calculated using the Spearman-Brown reliability formula (Winer, 1962).

One study of student-grading examined the confidence that the mean of the student scores differs significantly from the mean teacher scores (Davis & Rand,

1980). Although student's *t* test is a seemingly more sophisticated measure than agreement, comparing the means of scores awarded by teachers and by students has severe limitations. A teacher and a classroom of students can have identical means with very little agreement on individual grades.

Various measures of interrater reliability can be considered for comparing student-assigned grades with those of the teacher such as Cronbach's alpha, generalized Kappa, William's Index of Agreement, etc. (Abedi, 1996). However, we do not wish to measure interrater reliability, per se, but how well students' grades match those of the teacher. For these studies, the teacher's *grade assignment* is, by definition, the proper grade for a student. A student's grade and a teacher's grade are not of equal, a priori accuracy (this would be different if there were multiple teachers awarding grades). We are far less interested in the agreement between the grades that different students might award another's test and only in how well student grades match the teacher's grades.

For their meta-analysis, Falchikov and Goldfinch (2000) calculated the effect size between teacher- and student-awarded grades, converting agreement into effect size (ES). ES is the difference in subgroup means in units of the whole group's standard deviation of the score distribution. A commonly used statistic in meta-analysis, in this application ES characterizes the magnitude of the difference in mean grade between students and teachers on a particular assignment. In studies of educational interventions, ES often characterizes the size of the effect of an educational intervention between pre and post conditions. Measured in units of standard deviation, the magnitude of the ES has generally accepted ranges, with "large" effects considered 1.0 or above. Studies often include the pre–post ES of a control group to compare to that of the ES of the experimental group.

ES is used in an opposite fashion in the two meta-analyses. Small effect sizes are to be interpreted as teachers and students having made similar judgments. When effect sizes are large, teacher–student agreement decreases considerably. Hence, a small effect size is seen as evidence of agreement. One problem with effect sizes is that as long as the overall distributions of grades awarded by students and by teachers are similar overall (i.e., have the same mean grade), effect sizes will be small. This is troublesome because effect size is sensitive to the mean grade assigned by each group (say a grade of "B") and not to the agreement for the work of each individual student. There can be vast level of disagreement concerning actual grades on a case-by-case basis. For example, all tests graded "A" by the teacher could be graded "F" by students and all "F" tests graded by the teacher could receive a grade of "A" from students and still the effect size could be exactly the same as if each group agreed perfectly, as long as the mean grade and standard deviation was undisturbed. In the two meta-analyses, no statistical significance levels were calculated for ES. While allowing for comparisons between studies, ES manifests many of the same weaknesses of the agreement metric.

Although high agreement is preferable to fairly substitute student-grading for that of the teacher, nondiscrete models offer some advantages. Correlation compares the difference in grades on a case-by-case basis (using the squared differences in score) and ascertains the goodness-of-fit of a linear relationship. Grades are not only ordered but arguably follow a linear scale with the difference between grades being the same size. If treated as a nonlinear scale, simple means would not be a fair method of aggregating the grades on multiple assignments. With the assumption of a linear scale, Pearson correlation appears to be a useful measure for comparison of student to teacher ratings. By using correlation, the panoply of related statistical tools can also be called into action. Linear regression can be used to account for variation between classes. Linear models can aid in examining interactions (e.g., how boys and girls may grade differently when using different treatments). The weakness of correlation is that it is not sensitive to students and teachers using vastly different rating scales. Correlation will be high as long as one set of grades can be transformed into another through a linear transformation (i.e., using multiplicative or additive constants). Although a measure of "agreement" would be low for students who systematically score peers lower than the teacher, correlation measures the amount of variance explained by a functional model relating teacher grading and student-grading. For example, if students grade on an A through B scale and their teacher grades on a C through F scale, correlation can be high while agreement can be zero.

Correlation can measure the linearity of the relationship between teacher grading and student-grading. Using a Monte Carlo simulation of 10,000 40-student classrooms having varying levels of agreement with their teacher's grades of A, B, C, D, and F, we investigated the similarity of agreement to correlation (Figure 1). Grades are assigned with Gaussian probability between A and F (a mean of "C" and a standard deviation of one letter grade). Random numbers are added to

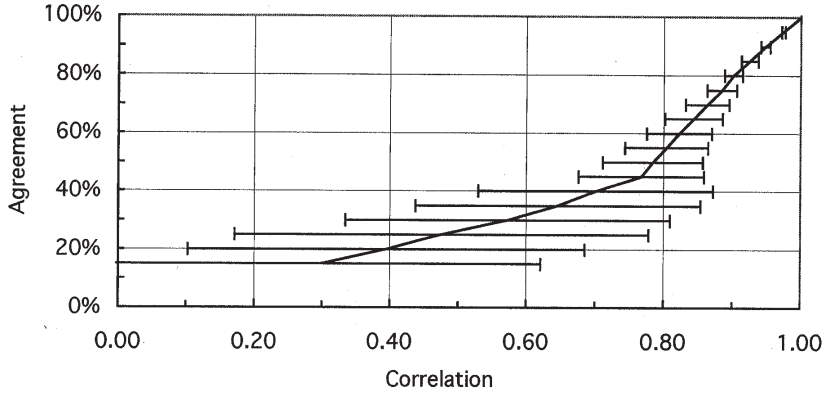


FIGURE 1 Monte Carlo simulation of grade agreement with correlation.

teacher grades to create a set of grades with added noise to simulate student scoring of the same papers. The noise added to these grades has an increasing range to simulate lowering agreement along with correlations. Agreement of letter grades varies considerably around particular correlation. Note that above an agreement level of 70%, agreement and correlation are very similar. Horizontal error bars are ±1 standard deviation about the mean. The slope of the graph in this region is approximately 2.0 with agreement decreasing 2% for a corresponding 1% decrease in correlation. When student grades assigned vary by more than one category (agreement < .60) the comparison becomes much more inaccurate. Agreement is cut off at 15% because such low agreements are rare if not a random function.

Other Findings in the Literature

The reliability of student-grading was higher when students were rating on achievement rather than on student effort (Boud, 1989). Weaker students tended to be more generous in grading themselves, contrary to the effect seen in stronger students (Boud, 1989). Varying views exist among educators concerning whether girls express less or equal confidence in grading their own papers (Baird & Northfield, 1992). Interrater reliability is increased by using rubrics with unambiguous scales and by using a small number of categories (five or fewer). A difference was found between the accuracy of grading of items that required lower order and higher order cognitive skills. Self-grading on lower order skills was well correlated with teacher's grades, but students had more difficulty grading items testing higher order understanding in a study of Greek and Israeli students (Zoller, Tsaparlis, et al., 1997). One study claimed a relationship between "undermarking" and "overmarking," with college freshmen being more generous with their own grades than college seniors (Falchikov, 1986).

Training and support in developing the skills necessary for accurate grading appears to pay off. When students were simply handed a rubric and asked to use it voluntarily, but were given no training in its use, they ignored the rubric (Fairbrother, Black, & Gil, 1995). When given the opportunity to self-assess using open-ended, qualitative statements instead of formal guidelines, students were terse or obscure to the point of uselessness (Baird & Northfield, 1992). One teacher found that it took a year for students to learn to use a rubric to accurately self-assess (Parkin & Richards, 1995). The most effective training exercise for grading involved initially reducing the cognitive complexity and emotional demands on students. Teachers lowered the pressure and increased the agreement by reducing the number of rubric categories so students only graded a few items in a training session. Alternatively teachers decided that the initial attempt at student-grading would not count for credit (Parkin & Richards, 1995; Weaver & Cotrell, 1986; Zola, 1992). One method involved pairs of students writing comments on each other's papers but not assigning grades initially as a first step, avoid-

ing the "culture shock" of students assuming responsibility for grading, which came later in the term (McLeod, 2001).

Asked to rate student behavior (i.e., behaviors that teachers find disruptive or bothersome) a high correlation was found between teacher- and peer-ratings of classmates. This was not true for a comparison between teacher- and self-ratings (Werdelin 1966). Five- through 7-year-old students rated their peers on who was better or worse than themselves at schoolwork (Crocker & Cheeseman, 1988). From these data, student rankings were generated and compared to rankings by the teacher. Correlations for peer-to-teacher rankings were higher (0.79) than for self-to-teacher rankings (0.64). When college students rated each other on their contribution to class discussions, students who consider themselves to have stronger skills in writing or test-taking expressed objections to being graded during discussions (Zola, 1992).

Janssen and Rijlaarsdam (1996) attempted to measure students' ability to apply the literature they studied to their own personal issues. When they found no measurable impact on such personal development, they ignored this finding by hypothesizing that the impact must occur over the long term. Studies of the impact of peer- and self-assessment have been found in the medical literature but not in the education literature. Twelve self- and peer-assessed health care providers in Mali were found to significantly differ in their compliance with government promulgated standards and the assessment of fever compared to a control group of 24 (Kelly, Kelly, Simpara, Ousmane, & Makinen, 2002). Improvement in the quality of sessions resulted from peer assessment and peer review in a study of family planning counselors in Indonesia (Kim, Putjuk, Basuki, & Kols, 2000) and in Ghana, Kenya, Nigeria, and Uganda (Lynam, Rabinovitz, & Shobowale, 1993).

Legal

The practice of peer-grading has risen to public attention quite dramatically because of legal actions begun by a concerned mother. Kristja Falvo initially complained to school counselors about the practice of classmates grading her own children's work and calling out these grades to the teacher. When Falvo asserted that this practice severely embarrassed her children (especially her learning disabled sixth-grade son, who scored a 47 on a quiz and was later teased), she was told that her children could always opt for having their grades privately reported to the teacher (Falvo v. Owasso Independent School District, 2000; Ragsdale, 2001). The school system felt this policy adequate and refused to enjoin teachers from having students trade papers to correct them.

Unsatisfied, Ms. Falvo pursed this matter. On October 6, 1998, she brought suit against the school district and three administrators in Federal District Court. The court decided in favor of the school system and let the practice of peer-grading stand, finding no violation of federal laws. Undeterred, Ms. Falvo then advanced

this matter to the Federal Appeals Court–10th Circuit on July 1, 2000, hoping to overturn the Federal District Court decision. The appeals court found the school system in violation of the Family Educational Rights and Privacy Act (FERPA), which prohibits schools from maintaining "a policy of permitting the release of educational records (or personally identifiable information contained therein …) of students without written consent of their parents" to anyone other than statutorily designated authorities or individuals (Falvo, 2000, p. 7). The penalty for violation of FERPA is the loss of federal funding to the offending institution. This federal ruling had the effect of prohibiting peer-grading in six states: Oklahoma, Kansas, Colorado, New Mexico, Utah, and Wyoming (Romano, 2001).

It may seem to the reader (as it did initially to the authors) that the primary legal issue involved in this matter is that of disclosing embarrassing low grades in the classroom and that peer-grading might be fine if these grades were simply reported privately. However, the case is really about the legality of peer-grading itself, as the U.S. Appeals Court states: "the prior act of one student grading another's paper itself constitutes a disclosure" (Falvo, 2000, p. 10). The Court of Appeals's view is that even if only one other student knows a student's grade, the law has been violated. The Bush administration, the National Education Association, and the American Federation of Teachers all disagree, arguing that only grades on an official transcript are protected by FERPA (2001), whereas the conservative Rutherford Institute (providing legal services and funding for the Falvo case) contends that even a test grade requires full privacy protection under federal law.

The Court of Appeals decision was itself appealed before the U.S. Supreme Court, which heard arguments on November 27, 2001. The appeal to the high court requests clarification on what are "educational records" under FERPA (Leslie, 2001). On February 19, 2002, the Supreme Court decided unanimously that peer grading does not violate FERPA.

During the Supreme Court arguments, Justice Stephen Breyer expressed concern that applying FERPA to peer-grading would stifle teachers' classroom practices (Walsh, 2001). In writing the *Opinion of the Court* released on February 19, 2002, Justice Kennedy agreed, first clarifying that the grades that students give each other do not constitute educational records that must be kept private (Owasso Independent School District No. I-011 v. Falvo, 2002). Making peer-grading illegal would produce a "drastic alteration" in the federal role in the traditionally local running of schools stating that "federal power would exercise minute control over specific teaching methods and instructional dynamics in classrooms throughout the country" (p. 9). The decision asserted that application of FERPA would hamstring teachers' freedom to pursue many commonly accepted activities that make distinctions between students, such as awarding "gold stars" and "smiley faces" on papers as a public disclosure of merit. Editorials have expanded on this to include topics ranging from the posting of exemplary work to the ubiquitous practice of recognizing honor roll students and valedictorians (Lehrer, 2001; Pearl, 2001).

In his decision, Justice Kennedy (Owasso Independent School District No. I-011 v. Falvo, 2002) described the benefits of peer grading, echoing those outlined in the research literature:

- Logistical—A teacher employing student-grading can spend more time "teaching and in preparation."
- Pedagogical—"It is a way to teach material again in a new context … . By explaining the answer to the class as students correct their papers, the teacher not only reinforces the lesson but also discovers whether the students have understood the material and are ready to move on."
- Metacognitive—"Correcting a classmate's work can be as much as a part of the assignment as taking the test itself."
- Affective—"It helps show students how to assist and respect fellow pupils." (pp. 7–8)

Research Questions

Motivated by the strong views of parents, teachers, and the courts, we designed our study to help address two issues of interest to teachers and researchers concerning student-grading:

1. *Student-grading as a substitute for teacher grades.* To what extent do students' grades of their own and others' papers reflect the grades of their teachers when using a grading rubric? Are student grades a valid substitute for teacher grades? Can they be used to reduce teacher time by ranking or sorting student grades into ordered groups? Are student grades more accurate under some conditions?
2. *Student-grading as a pedagogic tool for student learning.* To what extent does a student grading her own or others' papers improve her grade on the test if taken again? Do these effects differ by performance level of the student or other attributes?

Data Collection

This study was conducted in four seventh-grade, general science classrooms in the month of February. Each class was a heterogeneous section of the same course. Class means for prior tests and quizzes to date were very close, and were all within 0.5% of 85%. Each class experienced a different, randomly assigned intervention when it came to scoring tests taken as a normal part of studying a biology unit on the classification of organisms. These interventions were:

- control (teacher-only grading),
- self-grading (plus teacher grading), and
- peer-grading (plus teacher grading).

The test was generated by the teacher from a variety of sources. It included 9 fill-in-the-blank items, 7 classification tasks, 13 questions to be matched with answers on the test form, and 5 constructed response items, including an elaborate application of biological classification to the characters in Tolkein's *The Hobbit*. The test was designed as "open-note" with students able to employ their class notes while taking the test. The test contained easy-to-grade single-answer questions (worth 40% of the grade). The remainder of the items required "higher order thinking" application, analysis, and synthesis (Bloom & Krathwohl, 1956).

For students grading others' tests and for the teacher in this study, identification numbers were substituted for student names. Neither the teacher nor fellow students could discern in which class a test had been given or which student in a class had filled out the test. One full period (approximately 40 min) was devoted to creation of a rubric that students would later use to grade tests. Each question was discussed and the essential characteristics of a correct answer were listed. The total number of points for the question were distributed over these possible answers. After an initial rubric was constructed in each class, the teacher shared rubrics from the other classes and brought up any points that may have been missed. Generally students came up with the same complete answers; only the order of the attributes was different. An example of a rubric for one item, where each bullet is worth two points, follows.

Compare and contrast the classification systems of Aristotle and Linnaeus.
Similarity: used only observable characteristics of organisms;
Differences:

- Aristotle used where animals live (air, land, water) or plant size and structure;
- Linnaeus used body structure, color, ways of getting food;
- Linnaeus named using binomial nomenclature: genus-species in Latin;
- Linnaeus used many taxonomic levels: Kingdom, phylum or division, class, order, family, genus, species.

After the grading rubric was developed by each class, it was displayed during another class session when the tests were actually graded. These students had used self- and peer-grading extensively since the start of the year. This included developing rubrics and previously grading 10 quizzes and tests and approximately 40 homework assignments. Students were very familiar with the process of grading papers. Each item was corrected in turn. Students were encouraged to orally present answers from the test papers and compare them to the general rubric for constructed-response questions. Vigorous class discussions ensued regarding the accuracy and acceptability of particular answers and the amount of credit merited for each. Although the presenter did listen to his or her classmates' views, ultimately the decision for each test item rested with the grader. Students added up scores

for each item and gave an overall numerical grade. The teacher then graded the tests without knowing student identities or the student-assigned scores using the same rubric.

Approximately 1 week after the administration of the first test and grading, an identical test was administered under the same conditions as the first. Students in all four sections then peer-graded these tests using the same process as in the first test. The teacher once again blindly graded all these tests so that gains could be calculated. A week after the students had graded the second test, the teacher passed out a feedback form on which they could offer feedback as to their thoughts on the student-grading experience.

DATA

The data collected for this study are composed of 386 test grades. By grouping these test grades in different ways (e.g., matching student and teacher grades or pairing first and second tests by students) we address our research questions.

1. Class 1 consisted of 25 students. After the first test, each student corrected two tests from two other classes, one from class 3 and one from class 4. After the second test, each student corrected one test from another class. The second administration of the test was taken by only 24 students resulting in 24 matched pairs.
2. Class 2 consisted of 24 students. After the first test, each student corrected his or her own test. After the second test, each student corrected a test from another class. The second administration of the test was taken by only 22 students resulting in 22 matched pairs.
3. Class 3 consisted of 25 students. After the first test, students did not correct their own or peers' tests. After the second test, each student corrected a test from another class. The second administration of the test was taken by only 24 students resulting in 24 matched pairs.
4. Class 4 consisted of 27 students. After the first test, each student corrected one test from class 1. After the second test, each student corrected one test from another class. The second administration of the test was taken by only 25 students resulting in 25 matched pairs.

The teacher independently corrected each of the 101 first tests and 95 second tests. Data are summarized in Table 1.

ANALYSIS

Several statistical procedures were carried out on the dataset. Descriptive statistics on the various groups were calculated, establishing means and standard deviations

TABLE 1
Who Grades What? Counts of Grading in Four Classrooms

	First Test				
Class	*No One*	*Self*	*Peer*	*Second Test Peer*	*Matched Grades From Teacher*
1			50	24	24
2		24		22	22
3	25			24	24
4			27	25	25
Total	25	24	77	95	95

for each administration of the test. Using teacher-awarded grades only, gains from the first test to the second were calculated and a *t* test was used to establish the significance of the mean gain for each treatment group. Results are also reported as effect sizes (ES = Δ mean/standard deviation). Standard deviation is a pooled value for all the teacher-graded pretests. By class, they ranged from 12.7 to 16.8 (Table 2) and were the largest for the self-grading class, alerting us to control for pretest scores when building analysis of variance (ANOVA) models.

Correlations between peer-grades and teacher grades and self-grades and teacher grades were calculated showing the degree of similarity between the two. Overall differences in mean grade and patterns in grade based on student performance are reported, including training effects from the first to second administration of the test by using a control group.

Gains were calculated from the first to the second administration of the test for the various treatment groups and three levels of student performance (the upper, middle, and lower thirds based on grades awarded by the teacher for the first test).

In retrospect, it would have been wise to analyze our data to account for levels of teacher–student grading agreement for the different kinds of test questions and to calculate internal consistency (reliability using KR-20). We did not analyze and record these subscores at the time of grading. Because tests were eventually handed back to students, these data are not recoverable. Presumably where students exercised the most judgment for scoring open-ended items, agreement would be shown to be the lowest between teacher and students. Test–retest results of the two administrations of the test allowed calculation of a correlation coefficient for the temporal stability of scores. This reliability coefficient for teacher-awarded scores was 0.611 for the control group, 0.770 for the peer-grading group and 0.676 for the self-grading group. The reliability statistic can be affected by learning or memory of test items or the length of time between the two administrations.

With the potential that differences in gender and pretest scores between the treatment groups may be responsible for observed differences by treatment, *t* tests

TABLE 2
Comparison of Student- and Teacher-Graded Results by Treatment Group

Student-Graded		*Test 1*		*Test 2*				
Class	*Group Action*	*M*	*SD*	*M*	*SD*			
3	Control	71.56	12.73	72.50	13.88			
1,4	Peer-grading	69.05	14.02	71.87	14.01			
2	Self-grading	77.73	15.59	83.80	11.92			
	All groups	72.78	14.92	74.89	14.05			
Teacher-Graded		*Test 1*		*Test 2*		*Gain*		
Class	*Group Action*	*M*	*SD*	*M*	*SD*	*M*	*ES*	*t test, p =*
3	Control	72.50	12.73	75.02	13.36	2.52	0.17	0.302
1,4	Peer-grading	71.61	13.88	77.02	13.16	5.41	0.37	0.069
2	Self-grading	75.85	16.80	87.84	10.12	11.99	0.82	0.000
	All groups	72.91	14.54	79.02	13.39	6.12	0.42	0.000

and an ANOVA measured whether these natural variations were significant. Finally, regression models were built to find the most parsimonious set of variables that explain student gain based on treatment group (categorical), gender (categorical), and initial test score (continuous). Tests of normalcy of the dataset were also carried out. One missing element is that subscores on the tests by item type were not recorded.

Descriptive Statistics

Test means and standard deviations are reported in Table 2. Class means vary from 69.05% to 87.84% correct. Standard deviations vary from 10.12% to 16.80%. Gains have been calculated based on teacher-assigned grades in points and effect size. Gains are seen in the means of all three treatment groups and in the sample as a whole. Using a *t* test, only the self-grading group was found to have a pre–post gain significant at the $p \leq 0.05$ level.

Student-Teacher Grading Comparison

The grades that students awarded to themselves or to their peers are presented in Figure 2. Although peer-grading students tended to undergrade, self-grading students tended to overgrade.

Table 3 shows the various measures of agreement for peer-graded and self-graded tests when compared to teacher grades. Agreement measures the fraction of

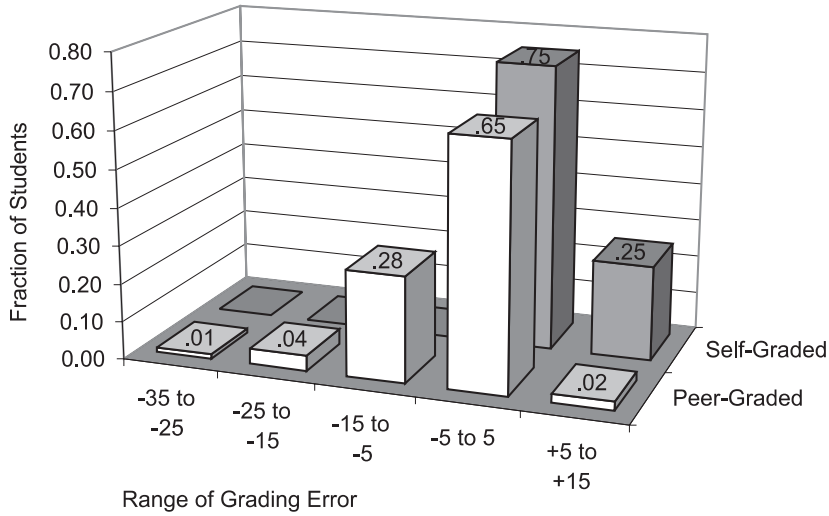


FIGURE 2 Comparison of tallied student grades showing the departure from teacher's grade.

student-teacher pairs that are in the same grade level (A, B, C, D, and F). The Kappa statistic corrects for chance matches of these pairs. The weighted Kappa statistics give "partial credit" for grade pairs that are close to each other and less credit if they are less similar using either a linear or quadratic weighting scheme. One can see that the quadratic weighted Kappa has a value approaching that of the correlation coefficient, as has been noted in other studies (Shuster, in press). Self-grades appear to show a highly correlated pattern with teacher grades r = 0.976) demonstrating high interrater reliability.

Figure 3 shows a scatterplot of the data, plotting student-assigned grades against the grades that were awarded by the teacher. There are many more peer-grades than self-grades because the second administration of the test was peer-graded by all students. The majority of self-grades are seen to be above the diagonal "perfect agreement" line, showing that students awarded slightly higher grades (+1.9 points) on average to themselves than their teacher assigned. Peer-grades are less highly correlated with teacher grades $r = 0.905$) and appear mostly

TABLE 3
Measures Comparing the Similarity of Teacher and Student Grades

Grading Comparison	*Agreement*	*Kappa*	*Weighted Kappa (Linear)*	*Weighted Kappa (Quad.)*	*Correlation*
Self to teacher	.79	.74	.89	.96	.976
Peer to teacher	.61	.51	.71	.85	.905

below the perfect agreement line. Peer-grades average 3.3 points below teacher grades with two substantial outliers attributed to peer-grading. There are also some additional substantial outliers. For test 1, the standard deviations for teacher and student grades are lower for peer-grading than for self-grading. There are a few more higher performing students in the self-grading group, which raises the standard deviation for scores judged by the students themselves and their teacher.

The high correlations reflect the degree to which the data-points approach the regression line, obscuring the differences between peer- and self-grading. Figure 4 compares the teacher's grades to students' by subtracting the teacher's grades from student-assigned grades, which can be considered the error in student assigned-

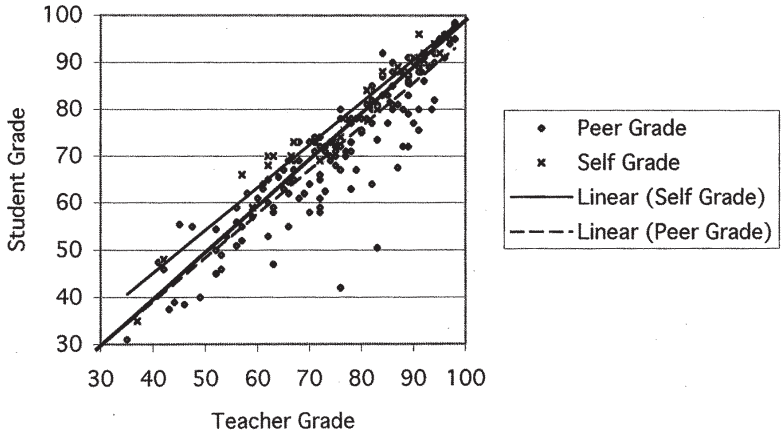


FIGURE 3 Comparison of teacher grade with peer- and self-grades and best-fit lines.

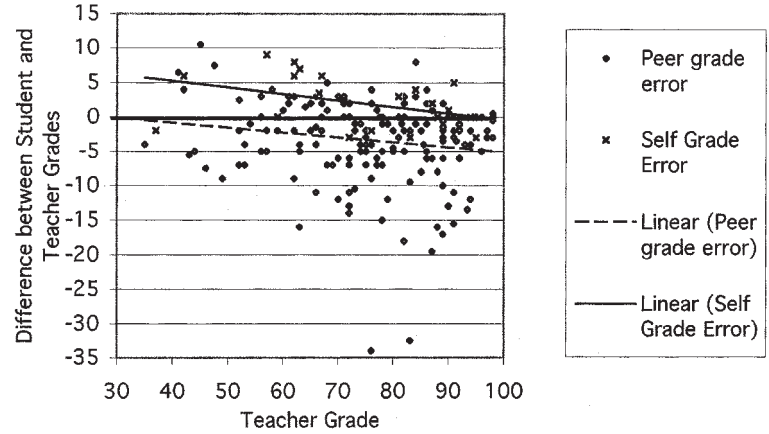


FIGURE 4 Error in student grading compared to teacher grades with best-fit lines.

grades. Here the trends are more apparent with self-grades uniformly averaging about 5 points higher than peer-grades. Poorly performing students tended to overrate themselves in comparison to teacher-assigned grades. Better performing students tended to be underrated by their peers.

Training Effects

Because the test was administered twice, student–teacher grade comparisons can be separated into two groups and compared to see whether student-grading accuracy improved. It can be argued that the process of learning to grade a peer's paper improves with practice. Using 73 matched sets of 2 peer-grades and 2 teacher grades for 2 administrations, we compared the difference between first and second peer-grading correlations with the teacher's grades. There is high correlation between teacher and student grades, with $r = 0.92$ for test 1 and $r = 0.91$ for test 2, barely any change at all. Agreement calculated between these two groups changed from 0.64 to 0.61, also a small change. This shows no training effect between these two grading opportunities. Perhaps this is because students were well trained by the time of the first test, having had extensive practice since the start of the school year. A second round of grading showed no evident improvement in grading skill, just a slight decline.

Differences Between Treatment Groups

The authors were concerned that the self-grading group, although selected at random, started with a somewhat higher mean teacher-awarded test score than the other groups. Perhaps this was responsible for much higher gains on the second test if the students had a higher ability. We approached this issue in three ways.

1. All groups were analyzed by assigning students into one of three achievement levels: lower, middle, and upper thirds (with breakpoints at 67 and 82 points).
2. An ANOVA was carried out to gauge whether there were significant differences in the treatment groups in regard to initial test grade and also gender distribution. No significant difference was found for either at the $p \leq 0.05$ level.
3. Three linear models were constructed using three treatment categories while accounting for the variables of initial test grade. Neither treatment group nor student sex was a significant predictor of initial test grade.

Table 3 includes results from a *t* test applied to first and second test score distributions by treatment group. Neither control nor peer-grading groups show significant gains in group means (at the $p \leq 0.05$ level). A study that drew on a larger sample of students might show significance. Yet even at this sample size, the magnitude of the gain in the self-graded groups is large, an effect size of 0.82 standard

TABLE 4
Analysis of Variance for First Test Grade

Source	*df*	*Sums of Squares*	M^2	*F Ratio*	*p*
Constant	1	504942.00	504942.00	2353.30	≤ 0.0001
Treatment group	2	305.39	152.69	0.71	0.4936
Gender	1	72.79	72.79	0.34	0.5617
Error	91	19525.90	214.57		
Total	94	19870.10			

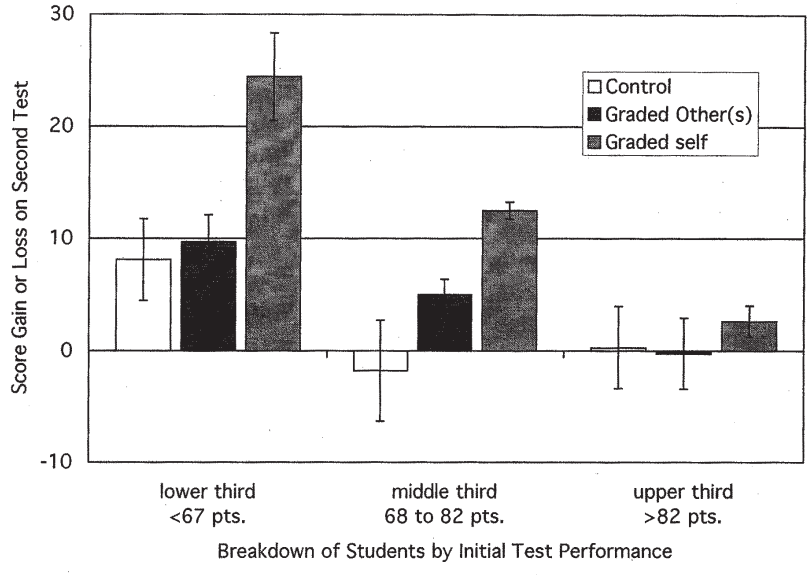
deviations. It is possible that considering the groups as a whole may mask more subtle effects only visible when considering gains for students at different achievement levels. For this reason we analyzed gain after breaking students into three groups based on initial scores on the first administration of the test. The lower third of the students had teacher-assigned grades below 67 points, the upper third scored above 82. The middle third scored between these two breakpoints.

An ANOVA of the initial test scored by the teacher shows that neither treatment group nor gender contributed significantly to differences in student test grades (Table 4). This helps in interpreting gains in test score in the second administration of the test.

Gains displayed by student achievement level on the initial test are shown in Figure 5. Here the error bars enclose one standard error of the mean, and each treatment group is shown by a differently shaded bar. Distance above or below the zero line shows the gain or decline for each performance group. Any condition-related effect was certainly not large enough to advance the lower third of the students (based on their pretest) above their higher performing peers (on their posttest). For the control condition, only the lower third of students gained significantly, about 10 points above the baseline. Neither the middle-level nor the upper-level students in the control group improved significantly from pre- to posttest. For students who graded their peers, the lower and middle thirds improved significantly, but the gain at the lower level is no higher than that for students in the control group. Upper-third students who graded their peers did not improve. For students who self-graded, gains are seen at all levels of achievement with those at lower levels gaining most. Gains for self-graders are well above the level of the control group. Small gains are seen at the highest achievement level where there is less room for improvement. Peer-grading appears slightly more effective than the control group, whereas self-grading is more effective at all performance levels.

ANOVA

ANOVA aids in data analysis through inclusion of variables that, along with differing treatment groups, can also affect outcomes. These variables represent alterna-



Note. Error bars enclose ± 1 SE of the mean gain.

FIGURE 5 Gain by student achievement levels for three different conditions.

tive hypotheses to the hypothesis that the treatments alone are responsible for gains in test scores. For ANOVA to be used effectively, the distribution in outcome variable must be Gaussian. The distribution in gain is roughly normal in shape, with a kurtosis of 1.012 and a small degree of skew of 0.137. One reason for this skewness is the asymmetry of distribution in grades. Although the test appears rather difficult, with a first administration mean of 72.88%, there is a smaller limit on how much more high-scoring students can achieve compared to those who score lower than the mean, resulting in a "ceiling effect."

With the second test used as a dependent variable, a linear model was built from categorical and continuous variables (Table 5). The variable first test grade is significant. Each point in the initial test grade predicts only .602 of a point in the second test grade, due to the "ceiling effect." Treatment group is also significant, with the second test grade of the control group being significantly lower than average and the self-grading group performing significantly higher. The Scheffe Post Hoc test is conservative in that it compares the predicted values of the second test score for the different treatment groups assuming that the likelihood of a Type 1 error is, at most, alpha ($p \leq 0.05$) for any of the possible comparisons between levels of treatment. The Scheffe test shows that the self-grading group significantly outperformed the peer-grading and control groups, which did not differ significantly

TABLE 5
Analysis of Variance Linear Model for Second Test Grade

Source	*df*	*Sums of Squares*	M^2	*F Ratio*	*p*
Constant	1	504942.00	504942.00	2353.30	≤ 0.0001
First test grade	1	7089.83	7089.83	86.37	≤ 0.0001
Treatment group	2	1482.94	741.47	9.03	0.0003
Error	91	7470.09	82.09		
Total	94	19870.10			
Results		*Coefficient*	*Standard Error*	*t Ratio*	*p*
Constant		35.82	4.85	7.38	≤ 0.0001
First test grade		0.60	0.06	9.29	≤ 0.0001
Treatment group					
Control		–4.40	1.46	–3.02	0.0033
Peer-grading		–1.95	1.25	–1.57	0.1209
Self-grading		6.35	1.50	4.23	≤ 0.0001
Scheffe Post Hoc Tests		*Difference*	*Standard Error*		*p*
Peer-grading–Control		2.45	2.26		0.5562
Self-grading–Control		10.76	2.68		0.0006
Self-grading–Peer-grading		8.30	2.34		0.0028
Treatment M		*Expected M*	*Actual M*		*Cell Count*
Control		75.26	75.02		24
Peer-grading		77.72	77.02		49
Self-grading		86.02	87.84		22

from each other. Adding student gender did not improve the model, nor did including two-way interactions between the variables. This model accounts for 62.4% of the variance in posttest scores.

A repeated measures ANOVA was carried out to account for the difference in teacher-awarded test scores while accounting for individual differences between students while nested in each treatment group (Table 6). The student identification number provides an additional 98 degrees of freedom and the model includes students who took only one of the two tests. With the test number (1 or 2) treated as a continuous variable, the second administration accounts for a significant gain of 6.57 points. Student number is also significant at the $p \leq 0.0001$ level. The two-way interaction between treatment group and test number is significant at the level $p = 0.0113$, showing that there is a difference in pre- and posttest scores by treatment when accounting for differences between students and the gain resulting from a second administration of the test. By adding to-

TABLE 6
Repeated Measure Analysis of Variance Linear Model for Test Grade

Source	*df*	*Sums of Squares*	M^2	*F Ratio*	*p*
Constant	1	1127920.00	1127920.00	3737.60	≤ 0.0001
Test #	1	1805.57	1805.57	31.39	≤ 0.0001
Student #	98	29573.80	301.77	5.25	≤ 0.0001
Treatment group	2	96.05	48.03	0.16	0.8531
Treatment group by test #	2	541.57	270.78	4.71	0.0113
Error	92	5291.54	57.52		
Total	195	39296.90			

Results	*Coefficient*	*Standard Error*	*t Ratio*	*p*
Constant	66.82	1.82	36.73	≤ 0.0001
Test #	6.57	1.17	5.60	≤ 0.0001
Treatment group				
Control	3.36	6.15	0.55	0.5859
Peer-grading	–0.49	5.22	–0.09	0.9258
Self-grading	–2.87	6.25	–0.46	0.6468
Treatment group by test #				
Control by test #	–4.05	1.72	–2.35	0.0210
Peer-grading by test #	–1.29	1.47	–0.88	0.3809
Self-grading by test #	5.34	1.77	3.03	0.0032

Treatment M	*Expected M*	*Actual Gain*	*Cell Count*
Control	75.22	75.02	49
Peer-grading	76.88	77.02	101
Self-grading	87.77	87.84	46

Treatment Gains	*Expected Gain*	*Actual Gain*
Control	5.88	2.52
Peer-grading	4.79	5.41
Self-grading	9.04	11.99

gether the effects of test number, treatment group, and treatment group by test number, one can calculate the expected gain based on this model. The expected gain for the control and peer-grading groups is similar to the observed gain, whereas that for the self-grading group is nearly double. One can see that the expected gains depart from the actual gains due to differences between individual students. Despite random assignment of treatments to classes, the peer-grading group was made up of more lower performing students (as judged by their pretest scores) than the control group, whereas the self-grading group was slightly

higher performing initially. The repeated measure model accounts for these differences as if the students were randomly assigned. This model accounts for 86.5% of the variance in test scores. Gender was not found to be a significant variable.

DISCUSSION

This study strove to identify ideal conditions for student-grading. Students were well trained in grading and had a rubric that codified their own and the teacher's judgment. Students could ask questions and work together during the teacher-supervised grading process; they were not rushed. For those who were peer-grading, names were hidden. Also, the teacher graded tests blind to student names, removing a potential source of bias. The accuracy that results from these preparations and precautions should be viewed as an ideal situation; it would probably be difficult to improve conditions further. Less training, no guidelines, and lack of blind review by peers would serve to reduce student–teacher agreement on grades and presumably any resultant student learning from the process.

Student-Grading Compared to Teacher's Grading

Can student-grading substitute for teacher grades? Undoubtedly this depends on whether a "student [can] come to hold a concept of quality roughly similar to that held by the teacher" (Sadler, 1989, p. 121). The very high correlations between grades assigned by students and those assigned by the teacher in the study support the idea that this kind of substitution can be attained even by seventh graders. Yet, because student-awarded grades are not identical to teacher grades, some sort of moderating procedure could be used to control for bias or distortion (Boud, 1989). Without students awarding exactly the same grades, a teacher is obligated to add some oversight to the process of student-grading.

Assignment of grades by students produces an ordering of test grades that can contribute to a reduction in effort that many teachers seek in their grading responsibilities. One can even argue that teachers will do a better job grading if they initially use student-grading, because the more information available to the teacher, the greater chance for accurate grading (Ory & Ryan, 1993; Reed, 1996).[1] In us-

[1]One of the reviewers of this article disagrees with this reasoning, noting: "Because the teacher has no idea who got mis-scored, it will take a greater effort to correct the discrepancies." Although this would be true if a large fraction of tests were out of order, after training student-grading is very similar to that of the teacher. The teacher must only look for gross anomalies and put them in their proper place. This means attending primarily to the lower performing self-graders and the higher scoring peer-graded students.

ing self-grading repeatedly, the teacher can, over time, identify those students who tend to "pad" their grades, awarding higher marks than they legitimately deserve. Teachers can act to correct these anomalies by informing students of their accuracy.

In our study, self-grading had higher agreement with teacher grades than peer-grading using several measures (Table 1). We found peer–teacher correlations of 0.905 and teacher–self correlations of 0.976. Self-grading has the potential of outperforming peer-grading, being a closer substitute for teacher grades in terms of correlation and also of agreement. Perhaps in our study this is because students knew the teacher was going to independently grade the papers and would be aware if they under- or overgraded. We speculate that students may be more conscientious in grading their own papers than those of a randomly selected, nameless peer, because they have more at stake in making sure their own paper gets graded accurately.

We must be careful not to interpret these high correlations as meaning that the test is a valid measure of student learning. High reliability is necessary, but is not a sufficient measure of test validity. The fact that students can accurately grade a test does not mean that the test is actually measuring what is intended (Sadler, 1989). One must also pay attention to the fact that there appears to be some systematic bias. Correlations between the teacher-assigned grades and those awarded by students show that lower performing students tended to overestimate their own grades in this study (as also found in Boud, 1989, and Parkin & Richards, 1995). Why? Lower performing students may not understand the rubric as well or may simply bow to the internal pressure to report better performance on their own tests. It may be that a lack of understanding of the rubric is the result of a lack of learning.

Do Students Learn From Grading Themselves and Others?

By using a second, unannounced administration of the test, this study measured (apparently for the first time in the research literature) the differences between peer- and self-grading. The literature hypothesizes that self- and peer-grading can provide feedback that a student can use to gain further understanding (Sadler, 1989). With 95 matched, pre–post pairs, significant and large gains are seen for the class that self-graded only. Peer-grading did not result in significant gains. The linear model controls for student scores on the initial test. The repeated measure model accounts for differences by individual student nested by treatment. Simply put, students do not appear to learn very much from grading the papers of others.

It is reasonable to expect that students would achieve higher scores on a second, identical administration of a test 1 week later. We controlled for this effect by measuring the gains of class 3 in which there was no treatment; students did not grade others' or their own papers between test administrations. For the control group, only the lowest level student group shows a statistically significant gain above zero

(Figure 5). Breaking down gains by ability level for the peer-grading group shows that lower and middle-level students appear to achieve a gain. But only for the middle-level group is the gain of peer-graders significantly larger than for the control group ($p \leq 0.05$). Students at all levels appear to benefit from self-grading, with significant gains at the lower and middle levels. The gains for lower and middle-level self-graders are significantly higher than for the control group and peer-graders. By measuring gain from a second administration of an identical test, if anything, we are probably overestimating the learning that may be taking place. The repeated measure model estimates that a second administration of the test boosts student test scores, but only self-grading increases gains well beyond that level. We speculate that the administration of another test with different questions would not attain this level of gain.

Our study did not seek to measure metacognitive learning, but students had a chance to reflect on the activity of grading others or themselves in writing just after the experience. Many felt it to be a valuable activity, after the initial oddness of the idea wore off. Zoller and Ben-Chaim (1998, p. 142) found that university students had a range of opinions about the value of self-grading, from negatives of "lack of objectivity, deep understanding and maturity on the part of the students" to qualified acceptance for objective test questions and the view that they can learn from their own mistakes. Comments were collected from students in our study a week after they had graded tests after the second administration. Statements included:

> At first the idea of grading my own test seemed wrong, that's the teacher's job. But after seeing the answer key [the rubric] and being able to talk about it, I understood some of it better. (Mary)

> This felt strange at first … but by the end … I realized grading was another way of learning. (Jane)

> Remembering not to put your name on [the test] was confusing [to keep the identity of students secret] … then it made sense. I think grading my own paper helped me more than grading someone else's. (John)

Implementation in the Classroom

How then might peer or self-grading be best implemented in the classroom? For optimal student-grading, we suggest training, blind grading, incentives for accuracy, and checks on accuracy compared to teacher grades.

There are a variety of forces on students that serve to complicate the introduction of student-grading. This kind of activity represents a change in role for the student, a break with the traditional authority structure in the classroom (Weaver & Cotrell, 1986). This can be troubling to some students, and for all it is better to gradually in-

troduce any novel aspect into a course. In self-grading, some students worry that other students might view them as boastful, should they give themselves high grades. This "overly modest" approach has been reported as affecting girls more than boys. One administrator stated:

> In my experience girls (and indeed female teachers) consistently undersell themselves. In a self assessment/partner assessment exercise … involving only pairs of friends the vast majority [of girls] over several years have marked themselves down on academic ability … (Baird & Northfield, 1992, p. 25)

We did not find this to be true, as another study disagrees:

> … most of our evidence led us to conclude that the undoubted tendency to underestimate achievement was shared roughly equally by boys and girls … (Baird & Northfield, 1992, p. 72)

Not knowing whose paper you are grading is not possible for those self-grading, but substituting numbers for names eliminates an accuracy-lowering distraction in peer-grading (McLeod, 2001). For peer-grading, this new role produces conflicting loyalties. Teachers must be aware that friendship or simply camaraderie with classmates can cloud the accuracy of peer-grades (Darling-Hammond et al., 1995).

Students may bridle at implementation of self- or peer-grading because they see grading as the exclusive purview of the teacher. They cite a lack of training, the fact that peers do not take the responsibility of grading others seriously, or that student-grading detracts from focusing on activities from which a student may feel she learns more (Weaver & Cotrell, 1986).

There are several ways in which student-grading can be the initial step in officially awarding students' grades. Although teachers often scrutinize for errors and improve rubrics that they employ themselves, a new role for teachers that wish to employ student-grading is making sure graders have used the rubrics properly (Weaver & Cotrell, 1986) or help to create them. Teachers can simply place in rank order the student-graded tests, with each test having its accompanying rubric. Teachers can then contribute their additional judgment to reorder tests and assign grades adding the benefit of their judgment. It remains to be seen whether time savings could result from this two-step process.

Caveats

This study was carried out in the four classrooms of a single teacher. We must be cautious in generalizing more broadly. Other schools, teachers, and classrooms might behave differently and yield different results. Yet, we sought to characterize

the highest possible student accuracy, as well as achieve the greatest possible gain. These kinds of ideal conditions are rare in school settings. Finding a high correlation between student- and teacher-awarded grades gives a glimpse of what is possible. The lack of significance and small magnitude of test gains for peer-graders may be troubling for those who promote its potential for student learning.

Future Efforts

This study was designed to answer its research questions, but as in all research, additional questions arose during analysis that could not be answered with the data collected. Were we to embark on another study or miraculously be able to begin this study anew, we would alter several facets.

Although the collected data settle the question of similarity of student grades when compared with the teacher's, these correlations would be easier to interpret if several teachers graded the students using the rubric. In addition, having the same teacher regrade students would put an upper limit on the highest level of rater agreement to be expected, because a teacher grading the same papers twice should produce the highest correlations.

The substitution of numbers for student names allowed the tests to be graded in a blind fashion by teacher and students. Presumably if the identities of students were not hidden, friendships between students and prior expectations of the teacher would lower the accuracy of grades. It would be interesting to quantify the difference in correlations between grades assigned when the grader knows the test taker and grades assigned when the student is unknown to his or her grader.

We carried out this study with four classes taught by a single teacher (the second author). Employing more teachers with their own classrooms would increase the number of students and allow better generalization. In addition, effects with smaller magnitudes can be identified with a larger number of subjects.

To measure student gain from grading papers, we relied on administering the same test a week later. Using an identical test likely sets an upper limit on possible gain. Tests on the same subject matter, but that employ different items (equivalent forms), would eliminate the possibility that students had just memorized the correct answers. Performance on two equivalent test forms would be a more valid measure of whether students had learned from the grading experience.

By raising students' regard for the accuracy of the grades they award, student agreement with teacher grades may increase even more. One way to do this is to give rewards to students for assigning grades closest to those of their teacher.

Most of the literature found is limited to college-level students. Are older, more mature students more accurate in their grading? Do they learn any more from self-grading? Our results suggest that middle school students have the capability for accurate grading and can learn much from the process.

Ours was a rather difficult exam, with many questions that required higher order thinking. It would be interesting to see whether student-teacher agreement increased with tests requiring less judgment to grade, such as with multiple-choice and fill-in-the-blank items. Also the calculation of subscores for different kinds of questions would allow us to measure aspects of the internal consistency of the test. A more extensive test-retest protocol (using correlations between scores) would help to establish the reliability of the test.

SUMMARY

Student-grading is best thought of not as an isolated educational practice, but as a part of the system of learning and assessment carried out in a teacher's classroom. It involves sharing with students some of the power traditionally held by the teacher, the power to grade. Such power should not be exercised recklessly or unchecked. Such power in their own hands or in those of their peers can make students uncomfortable or wary. Teachers should train their students in the skills entailed in accurate grading and should monitor students for accuracy in self- or peer-grading. When used responsibly student-grading can be highly accurate and reliable, saving teachers' time. *In this study, self-grading appears to further student understanding of the subject matter being taught.*

This study shows that high levels of interrater agreement are possible between students and teacher when students grade their own or others' papers. Teachers should be aware that lower performing students tend to inflate their own grades, whereas the grades of higher performing students may suffer when graded by others. Neither of these tendencies will change the rank-order of papers. Teachers should monitor the accuracy of student grades and not simply assume that grades are awarded fairly by all students (Boud, 1989). *Students should be trained to grade accurately and be rewarded for doing so.*

There are legitimate reasons for questioning the value of student-grading, especially because statistically rigorous research in this area appears rare. If implemented poorly, the grades can be unreliable and students may not learn anything valuable from the process. The U.S. Supreme Court supports the use of student-grading on legal grounds and, unlike the lower courts, expressed concern with the impact of such teaching methods on furthering student understanding. Self-grading will better maintain privacy of student grades, should that be a concern of classroom teachers, school administrators, and parents. With peer-grading, at least one other student is aware of a peer's grade if names are left on the paper. Blind grading protects privacy fully for those teachers who use peer-grading. It is doubtful whether Ms. Falvo would have sued Owasso School System if blind peer-grad-

ing had been used in her son's classroom but he probably would have learned more if he were simply allowed to grade his own test.

The U.S. Supreme Court decided not to intervene in the choice of teaching and assessment methods of the nation's teachers. Many students share their grades voluntarily with others, whether or not teachers actively try to protect their privacy. With the simple, pragmatic cures of blind grading or self-grading at a teacher's disposal, privacy can be protected and a teacher can benefit from having students help save them time and perhaps learn something in the process. This study shows that student-grading can benefit teacher and student when carried out with care.

ACKNOWLEDGMENTS

Support for this research was provided by the Harvard University Graduate School of Education, Harvard College Observatory, and the Smithsonian Astrophysical Observatory.

The authors wish to acknowledge the help received from many sources. Judith Peritz and Maria McEachern provided assistance in library research, finding relevant journal articles. Cynthia Crockett, Julie Lebarkin, Zahra Hazari, Nancy Cook Smith, and Harold Coyle provided critical readings. Finally we wish to thank the students whose efforts in grading their own and others' papers helped us and will help teachers they have never met.

REFERENCES

Abedi, J. (1996). Interrater/test reliability system (ITRS). *Multivariate Behavioral Research, 31,* 409–417.

Agresti, A., & Findlay, B. (1997). *Statistical methods for the social sciences*. San Francisco: Prentice Hall.

Alexander, P. A., Schallert, D. I., & Hare, V. C. (1991). Coming to terms. How researchers in learning and literacy talk about knowledge. *Review of Educational Research, 61,* 315–343.

Baird, J. R., & Northfield, J. R. (1992). *Learning from the PEEL experience*. Melbourne, Australia: Monash University.

Black, P., & Atkin, J. M. (1996). *Changing the subject. Innovations in science, math, and technology education*. London: Routledge.

Black, P., & Harrison, C. (2001). Self- and peer-assessment and taking responsibility, the science student's role in formative assessment. *School Science Review, 83,* 43–48.

Bloom, B. S. (1971). Mastery learning. In J. H. Block (Ed.), *Mastery learning, theory and practice* (pp. 47–63). New York: Holt, Rinehart, and Winston.

Bloom, B. S., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives. The classification of educational goals, by a committee of college and university examiners. Handbook I, Cognitive domain.* New York: Longmans, Green.

Boud, D. (1989). The role of self-assessment in student grading. *Assessment and Evaluation in Higher Education, 14,* 20–30.

Brown, A. (1987). Metacognition, executive control, self-regulation, and other more mysterious mechanisms. In F. E. Weinhart & R. H. Kluwe (Eds.), *Metacognition, motivation, and understanding* (pp. 65–116). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Burke, R. J. (1969). Some preliminary data on the use of self-evaluations and peer ratings in assigning university grades. *Journal of Educational Research, 62,* 444–448.

Crocker, A. C., & Cheeseman, R. G. (1988). The ability of young children to rank themselves for academic ability. *Educational Studies, 14,* 105–110.

Darling-Hammond, L., Ancess, J., & Faulk, B. (1995). *Authentic assessment in action: Studies of schools and students at work*. New York: Teachers College Press.

Davis, J. K., & Rand, D. C. (1980). Self-grading versus instructor grading. *Journal of Educational Research, 73,* 207–211.

Doyle, R. D., & Green, R. H. (1994). Self and peer appraisal in higher education. *Higher Education, 28,* 241–264.

Fairbrother, R., Black, P., & Gill, P. (Eds.). (1995). *Teachers assessing pupils. Lessons from science classrooms*. Hatfield, UK: Association for Science Education.

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

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

Falchikov N., & Goldfinch, J. (2000). Student peer assessment in higher education. A meta-analysis comparing peer and teacher marks. *Review of Educational Research, 70,* 287–322.

Falvo v. Owasso Independent School District (2000), Court of Appeals, 10th. U.S. Circuit.

Family Educational Rights and Privacy Act, 20 U.S.C. 1232g; 34 CFR Part 99.

Fleiss, J. L., & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement, 33,* 613–619.

Janssen, T., & Rijlaarsdam, G. (1996). Students as self-assessors, learning experiences of literature teaching in secondary schools. In E. Marum (Ed.), *Children and books in the modern world. Contemporary perspectives on literacy* (pp. 98–115). London: Falmer Press.

Kelly, E., Kelly, A., Simpara, C., Ousmane, S., & Makinen, M. (2002). *The impact of self-assessment provider performance in Mali.* Bethesda, MD: ABT Associates.

Kim, Y., Putjuk, F., Basuki, E., & Kols, A. (2000). Self-assessment and peer review: Improving Indonesian service providers' communication with clients. *International Family Planning Perspectives, 26,* 4–12.

Lehrer, J. (2001, October 28). *Courting student rights*. McNeil-Lehrer Productions. Public Broadcasting System. Retrieved October 25, 2005, from http://www.pbs.org/newshour/extra/features/july-dec01/sc_cases.html

Leslie, G. P. (2001). *Amicus Brief to the U.S. Supreme Court*, No. 00–1073, Owasso Independent School District v. Kristja J. Falvo.

Lynam, P., Rabinovitz, L., & Shobowale, M. (1993). Using self-assessment to improve the quality of family planning clinic services. *Studies in Family Planning, 24,* 252–260

McLeod, A. (2001). *In lieu of tests*. Retrieved August 4, 2005, from National Teaching and Learning Forum's Frequently Asked Questions at www.ntlf.com/html/lib/faq/al-ntlf.htm

Neukom, J. R. (2000). *Alternative assessment, rubrics—Students' self assessment process.* Unpublished master's thesis, Pacific Lutheran University, Tacoma, WA.

Ory, J., & Ryan, K. (1993). *Tips for improving testing and grading*. Newbury Park, CA: Sage Publications.

Owasso Independent School District No. I-011 v. Falvo (2002), 534 U.S. 426.[Supreme Court of the United States. No. 00–1073. Decided 2/9/2003.]

Parkin, C., & Richards, N. (1995). Introducing formative assessment at KS3, and attempt using pupil self-assessment. In R. Fairbrother, P. J. Black, & P. Gill (Eds.), *Teachers assessing pupils. Lessons from science classrooms* (pp. 13–28). Hatfield, UK: Association for Science Education.

Pearl, N. (2001, August). Making the grade public. *On the Docket*. Medill News Service, Northwestern University, Medill School of Journalism. Retrieved August 6, 2005, from www.medill.northwestern.edu/archives/000035.php

Pfeifer, J. K. (1981). *The effects of peer evaluation and personality on writing anxiety and writing performance in college freshmen.* Unpublished master's thesis, Texas Tech University, Lubbock, TX.

Phelps, R. P. (1997). The extent and character of system-wide student testing in the United States. *Educational Assessment, 4,* 89–121.

Ragsdale, S. (2001, December 9). Ask the kids how peer grading feels. *Des Moines Register.* Retrieved October 25, 2005, from http://desmoinesregister.com/news/stories/c5917686/16663282.html

Reed, D. E. (1996). *High school teachers' grading practices: A description of methods for collection and aggregation of grade information in three schools.* Unpublished doctoral dissertation, University of California, Riverside.

Romano, L. (2001, October 20). Supreme Court to test Okla. schools' policy of pupils grading peers. *Washington Post,* p. A03.

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

Schuster, C. (in press). A note on the interpretation of weighted kappa and its relations to other rater agreement statistics for metric scales. *Educational and Psychological Measurement.*

Walsh, M. (2001). Grading case takes high court back to school. *Education Week, 21*(14), 28, 31.

Weaver, R. L., & Cotrell, H. W. (1986). Peer evaluation: A case study. *Innovative Higher Education, 11,* 25–39.

Werdelin, I. (1966). Teacher ratings, peer ratings, and self ratings of behavior in school. *Educational and Psychological Interactions, 11,* 1–21.

Winer, B. J. (1962). *Statistical principles in experimental design.* New York: McGraw-Hill.

Zola, J. (1992). Scored discussions. *Social Education, 56,* 121–125.

Zoller, U. (1993) Are lecture and learning compatible? *Journal of Chemical Education, 70, 195–197.*

Zoller, U., & Ben-Chaim, D. (1998). Student self-assessment in HOCS science examinations: Is there a problem? *Journal of Science Education and Technology, 7,* 135–147.

Zoller, U., Ben-Chaim, D., & Kamm, S. D. (1997). Examination-type preference of college students and their faculty in Israel and USA: A comparative study. *School Science and Mathematics, 97*(1), 3–12.

Zoller, U., Tsaparlis, G., Fastow, M., & Lubezky, A. (1997). Student self-assessment of higher-order cognitive skills in college science teaching. *Journal of College Science Teaching, 27,* 99–101.