Journal of Experimental Psychology: Applied
2009, Vol. 15, No. 4, 319–333

1076-898X/09/$12.00 DOI: 10.1037/a0017841

Effects of Differential Feedback on Students' Examination Performance

Anastasiya A. Lipnevich
Educational Testing Service, Princeton, NJ

Jeffrey K. Smith
University of Otago

The effects of feedback on performance and factors associated with it were examined in a large introductory psychology course. The experiment involved college students ($N = 464$) working on an essay examination under 3 conditions: no feedback, detailed feedback that was perceived by participants to be provided by the course instructor, and detailed feedback that was perceived by participants to be computer generated. Additionally, these conditions were crossed with factors of grade (receiving a numerical grade or not) and praise (receiving a statement of praise or not). The task under consideration was a single-question essay examination administered at the beginning of the course. Detailed feedback on the essay, specific to individual's work, was found to be strongly related to student improvement in essay scores, with the influence of grades and praise being more complex. Generally, receipt of a tentative grade depressed performance, although this effect was ameliorated if accompanied by a statement of praise. Overall, detailed, descriptive feedback was found to be most effective when given alone, unaccompanied by grades or praise. It was also found that the perceived source of the feedback (the computer or the instructor) had little impact on the results. These findings are consistent with the research literature showing that descriptive feedback, which conveys information on how one performs the task and details ways to overcome difficulties, is far more effective than evaluative feedback, which simply informs students about how well they did.

Keywords: formative assessment, effects of feedback, grades, praise, college students

Supplemental materials: http://dx.doi.org/10.1037/a0017841.supp

Students in university courses typically receive one or more of three types of responses to the work that they produce: a grade, a statement of praise or concern, and some level of feedback on the specifics related to their performance (Orrell, 2006). The response that students receive often serves as a summary of their performance and provides information on how they can improve. These two different functions of the response are known as *summative* and *formative* functions of assessment (Scriven, 1967). The use of formative assessment to enhance student achievement has undergone a renaissance in recent years, leading to a variety of studies examining aspects of the relationship between formative assessment and students' ability to profit academically from such assessment (Schute, 2007; Symonds, 2004; Wiliam & Thompson, 2007). The formative function of assessment in university courses is the focus of this research.

Black and Wiliam (1998) proposed that the core of formative assessment comprises two types of information: (a) learners' current knowledge set and (b) the desired knowledge set. The discrepancy between the two represents a gap that is to be closed by the learner (Black & Wiliam, 1998; Ramaprasad, 1983). In order for assessment to facilitate learning, students need to receive information about the discrepancy between the actual and the desired state and effectively process that information. This information is commonly referred to as feedback (Ilgen & Davis, 2000; Kluger & DeNisi, 1996), and formative assessment can be conceptualized as a process through which learners receive feedback. However, not all feedback is the same, and not all feedback is equally effective in promoting learning (Hattie & Timperley, 2007; Kluger & DeNisi, 1996). The action taken by a learner in response to feedback depends heavily on the nature of the message, the way in which it was received, and the working contexts in which that action may be carried out (Black & Wiliam, 1998).

Effects of Feedback

Three comprehensive meta-analyses have been conducted over the past 20 years on the effects of feedback on achievement. Bangert-Drowns, Kulik, and Morgan (1991) found that although feedback was positively related to greater achievement in most settings, there was wide variability of feedback effects on performance. Overall, they concluded that the key feature in effective use of feedback is that it must encourage "mindfulness" in students' responses to the feedback. Kluger and DeNisi's (1996) meta-analysis demonstrated that although feedback typically improved performance, in one third of cases, presentation of feedback resulted in decreased performance. They contended that when feedback was accompanied by praise or critical judgments, the effectiveness of the feedback decreased and that feedback that showed participants how to reach correct solutions was more effective than were simple judgments of right or wrong responses.

Similarly, Hattie and Timperley's (2007) analysis found substantial variability in the effects of feedback. They reported that feedback about a particular task and how to do it is more effective than

Anastasiya A. Lipnevich, Educational Testing Service, Princeton, New Jersey; and Jeffrey K. Smith, Department of Education, University of Otago, Dunedin, New Zealand.

Correspondence concerning this article should be addressed to Anastasiya A. Lipnevich, Educational Testing Service, Rosedale Road, R-18, Princeton, NJ 08551. E-mail: a.lipnevich@gmail.com

feedback that focuses on praise or on punishments and rewards. Hattie and Timperly (2007) emphasized that feedback needs to address the questions of what the goals are, where the student currently stands in relation to those goals, and what the next steps should be for reaching the goals. They also noted that feedback focused on the level of the task, the processes required to complete the task, and self-regulatory task-related activities are more effective than is feedback focused on the person (typically, praise). Finally, Hattie and Timperly (2007) argued that feedback has "to prompt active information processing on the part of the learners" (p. 104).

We argue that the key to understanding the effects of feedback as it occurs via formative assessment in formal learning settings has to do with what Bangert-Drowns et al. (1991) call the *mindfulness* with which it is received or what Hattie and Timperley (2007) call *actively processing the information*. Unless students successfully process the feedback that they receive, there is little reason to believe that the feedback will have a positive effect on learning. But most research on feedback links feedback directly to subsequent achievement without considering the degree to which the feedback is successfully interpreted and processed. In this research, we examine how students in a university introductory psychology course use feedback on an essay exam to improve their work. This allows us to assess the degree to which the feedback was used to improve performance under different conditions, representing a tighter link between the intervention and the outcome.

Computer as Source of Feedback

Feedback has to have an origin. In some aspects of life, this origin may be inanimate, such as the direction and distance of a golf ball's flight after having been struck by a golf club. Other times, the source is an individual, typically a teacher. Advances in technology allow for feedback to come from a computer rather than from a teacher. The ability of a computer to provide feedback on objective assessments has existed for some time, but more recent advances have allowed for computer-based scoring of essays that include individualized feedback (Attali, 2004; Attali & Burstein, 2006; Landauer, Latham, & Foltz, 2003). The potential instructional benefit of computer-based feedback on such a labor-intensive task as marking essays is clear, but a serious question arises as to whether such feedback will be taken seriously by students.

One perspective on the question is that individuals will see computer-based feedback as essentially neutral (like the flight of a golf ball), but also not particularly accurate or helpful (Kluger & DeNisi, 1996; Lepper, Woolverton, Mumme, & Gurtner, 1993). A second perspective views computers as social actors (Nass, Moon, & Carney, 1999), with people attributing human characteristics to computers (Ferdig & Mishra, 2004; Nass, Fogg, & Moon, 1996; Nass, Moon, & Green, 1997). According to this perspective, students will respond to computer-provided feedback in the same way that they respond to human-provided information. Kluger and DeNisi (1996) summarized findings (albeit sparse) on computer versus instructor feedback and concluded that computer feedback was perceived as more helpful and accurate because of its tendency to bypass issues of attitude, affect, and stereotypes that are characteristic of human interactions.

Our goal in comparing instructor-based feedback with computer-based feedback is to provide information that directly speaks to this issue and comes from an experimental intervention. Although the research base at this point is not strong enough to make definite statements about how computer-based essay feedback will be received, it is our anticipation that participants will not take such feedback at a personal level and therefore will not have a negative reaction to it.

Grades as a Component of Feedback

The most common type of feedback that students receive is a grade, often with little or no additional commentary (Marzano, 2000; Oosterhof, 2001). Grades provide a convenient summary of student performance (Airasian, 1994), but how do grades perform in terms of a formative function? One of the main conclusions that Black and Wiliam (1998) drew from their review of literature on formative assessment is that descriptive feedback, rather than letter grades or scores, leads to the highest improvements in performance. Moreover, several studies have suggested that grades are actively detrimental and may hinder students' performance. For example, Butler and Nisan (1986) found that grades emphasized quantitative aspects of learning, depressed creativity, fostered fear of failure, and weakened students' interest. Butler (1988) found that students receiving comments specifically tailored to their performance resulted in a significant increase in scores on a task. Students receiving only grades showed a significant decline in scores, as did a group that received both grades and comments.

Interestingly, high achievers in all three feedback conditions sustained a high level of interest, whereas low achievers in the graded groups evidenced dramatic declines (Butler, 1988). It seemed that the presentation of a grade was particularly disconcerting when it indicated that performance was in some sense inadequate. The impact of receiving a grade may well depend on whether this grade is fundamentally good news or bad news (Black & Wiliam, 1998). It may be the case that no news is much better than bad news. The idea that the feedback delivered in different ways might have differential impact on students of different abilities has not been extensively studied. The design of the present study allowed for a critical examination of this issue by taking students' scores on initial drafts of their essays and splitting the sample into three subsamples based on those scores.

Explanations for the negative effects of grades on students' performance vary. Butler and Nissan (1986) and Butler (1988) proposed that grades inform students about proficiency in relation to others, whereas individualized comments create standards for self-evaluation specific for the task. They posited that even if feedback comments are helpful for students' work, their effect can be undermined by the negative motivational effects of giving grades and scores (Butler, 1988). Hattie and Temperley's (2007) model would see this as focusing the student at the level of the self rather than on the task or the processes that produced the performance on the task.

The empirical base for these arguments is not uniformly consistent. Although Butler's (1988) research found a negative effect of grades, Smith and Gorard (2005) found that students receiving grades and comments on their work outperformed students that received comments only. The research on the influence of grades is inconclusive, especially with the university students. Because most university assessment practices involve the assignation of grades, it is particularly important to investigate the impact of grades on student use of feedback information. We hypothesize that the presence of grades on

assessments at the university level has a negative impact on students' productive utilization of assessment feedback. The design of the current study allowed for a direct investigation into this issue, as well as into the question of how grades work in combination with praise and source of the feedback.

Praise as a Component of Feedback

Praise has been defined as "favorable interpersonal feedback" (Baumeister, Hutton, & Cairns, 1990, p. 131) or "positive evaluations made by a person of another's products, performances or attributes" (Kanouse, Gumpert, & Canavan-Gumpert, 1981, p. 98). Meta-analytic studies examining the effects of praise on motivation have shown that positive statements have a tendency to increase motivation across a variety of dependent measures (Cameron & Pierce, 1994; Deci, Koestner, & Ryan, 1999). This effect is not always strong, varies for different age groups, and often has been derived in the course of methodologically flawed studies (Henderlong & Lepper, 2002; Lepper, Henderlong, & Gingras, 1999).

The literature also includes examples of the negative impact of praise on students' learning. Baumeister et al. (1990) presented evidence that praise can both impede and facilitate performance. They argued that when praise focuses attention on the self as opposed to the task, cognitive resources are directed toward the self and not the task, hindering performance on more cognitively complex tasks. This argument is consistent with Hattie and Temperley's (2007) and Klueger and DeNisi's (1996) position that feedback focused on the self is not productive. We include praise as a factor in the design of the study, allowing for a direct investigation of the effects of praise, both in isolation and in combination with the factors of source of feedback and grades. We anticipate that praise will negatively influence students' performance.

Examining the Affective Outcomes of Formative Assessment Feedback

We have argued thus far the following:

- that feedback holds the potential to positively influence learning by prompting active involvement with the material to be learned;
- that feedback coming from a computer may be viewed differently from feedback coming from a course instructor;
- that including grades as a component of feedback may have a negative influence on how feedback is received by students;
- that the effects of feedback may be different for students of differing levels of initial achievement on a task;
- that including praise as a component of feedback may have a negative influence on how feedback is received by students; and
- that investigating these influences might more effectively be studied by relating them to the productive use of feedback by students than by the gains in learning.

Research in formative assessment frequently uses affective variables (such as mood, motivation, and self-efficacy) to explain the reactions that individuals have toward different feedback conditions (see, e.g., Butler, 1987; Butler & Nisan, 1987; Ilies & Judge, 2005). But there is little empirical research that actually examines how feedback influences affective response. For example, does the receipt of a grade actually result in a negative mood being induced, or a decrease in self-efficacy? We do not propose to explicate the exact nature of the workings of affective variables as moderator variables in this research, although we believe that this would be an excellent theoretical development. Instead, we propose to provide baseline information on whether differential feedback conditions actually result in differential affective responses in an experimentally controlled setting. Thus, although we believe that affective variables function as moderator variables in the feedback–response process, we use them as dependent variables here so that we can directly address the issue of whether they are influenced by different feedback conditions.

Several studies have shown that feedback containing praise leads to increased motivation (Delin & Baumeister, 1994; Ilies & Judge, 2005). Heniderlong and Lepper (2002) argued that favorable feedback cues would motivate children to work hard to sustain the approval of the evaluator, but that such behavior was transient, fading when the evaluator was no longer present. If praise is hypothesized to elicit positive affect, grades are often thought to lead to negative affect. Kluger, Lewinsohn, and Aiello (1994) argued that feedback received by individuals gets cognitively evaluated with respect to potential benefit or harm and for the need to take an action. More often than not, frustration, or other negative affective responses, followed by a sense of helplessness, prevents students from effectively carrying out a task and succeeding on it.

The negative effect of grades on students' performance can also be explained through the influence on students' self-efficacy. Generally, *self-efficacy*, or beliefs about one's competence, is known to be influenced by prior outcomes (Bandura & Locke, 2003; Vancouver, More, & Yoder, 2008). Although self-efficacy is typically conceptualized as a causal factor in educational and psychological research (Boekaerts, Maes, & Karoly, 2005; Vancouver, Thompson, & Williams, 2001), it is reasonable to consider it as an outcome of receiving feedback. A grade that causes students to question their sense of efficacy has the potential to negatively affect performance or to spur students to increased effort. Although there is evidence of the influence of feedback on motivation, mood, and self-efficacy beliefs, the research base is not extensive. We include measures of these three variables in the design as dependent variables to examine the degree to which they hold potential to help understand differences seen in the degree to which student work improves as a result of differential feedback.

Summary and Aims of the Present Research

The purpose of the study presented here was to systematically examine how feedback is received and used by university students under different conditions. We did this by investigating students' productive use of various forms of feedback. There have been a number of studies focused on aspects of grades, praise, and other feedback practices in higher education, but none that look specifically at all three of these aspects in combination, with experimental control, in a course setting in which the grades count for the

students. Because the consequence of the assessment for students is known to affect student performance (Wise & DeMars, 2005; Wolf, Smith, & Birnbaum, 1995), conducting the study as part of the grading system within a university course greatly adds to the ecological validity and generalizability of the findings. With the advent of computer-based essay scoring, it is now possible to provide computer-based feedback to students regarding their efforts. However, there is little to no research on how students react to feedback coming from the computer as opposed that coming from the professor in the course. Finally, there is not extensive research of issues related to students' affective response to detailed feedback, praise, and grades. In this research, we sought to examine the relationship between the feedback that students receive and their sense of self-efficacy, motivation, and mood.

Additionally, most prior research links feedback to ultimate learning, as opposed to successful engagement with the task and the feedback presented. This study examines improvement in performance on a specific, complex task through the productive engagement with feedback delivered in differing conditions. The design of the study also allows us to estimate the magnitude of the influence of various conditions of feedback and at differing levels of initial performance on the task studied in a fashion similar to that of Butler (1988). Finally, we are able to examine the interactions among the independent variables.

Method

In this experimental study, we investigated what happened when students were given the opportunity to revise an essay examination in an introductory psychology course on the basis of the receipt (or lack of receipt) of feedback on their first efforts. We also systematically varied whether students were told that the feedback came from the professor or from a computer essay-scoring program, whether students received a tentative, preliminary grade on their work, and whether they received a statement of praise and encouragement. This allowed us to study how important aspects of feedback influenced participants' subsequent behavior in their efforts to improve their work. The basic design of the study was a $3 \times 2 \times 2$ analysis of covariance: 3 levels of feedback (no feedback, feedback perceived to be from the professor, or feedback perceived to be from the computer program) $\times$ 2 levels of preliminary grade (presence or absence) $\times$ 2 levels of praise (presence or absence). The primary dependent measure was the score on the revised essay examination, and the covariate was the score on the first draft of the examination. In addition to using the examination scores, we used motivation, sense of self-efficacy, positive and negative affect, and perceived accuracy and helpfulness of feedback as additional outcome measures. These additional measures helped us to understand how the independent variables influenced improvement in performance.

Participants

We conducted the study with university students for several reasons. First, the format of a large, introductory university class taught by a single instructor allowed for the statistical power we were after without the confounding effects of multiple instructors. Second, we felt that university students were old enough and had enough experience in assessment settings to effectively process the information with which they were provided in the study. Third, the posttest measures we wanted to use could be administered to students at this level. Finally, there is a void in the research literature concerning this kind of controlled investigation with university students. Although there is ample research using university students in general, the use of formative assessment is not widespread in university level courses, and as a result, there is hardly any research on the topic at this level.

Participants for the experiment were students enrolled in introductory psychology courses at two public northeastern universities taught by the same instructor. The sample size for the experiment was 464 students, with 409 students attending University 1, and 55 students attending University 2. Separate analyses were run for the two samples to compare the distributions of key variables (i.e., the essay scores and affective measures) included in the current study; these variables were distributed normally for both samples, with nearly identical means and standard deviations. There were no differences in the basic findings of the study for the two different samples; therefore, the samples were merged.

The participants ranged in age from 17 to 51 years, with a mean age of 18.9 years ($SD = 2.5$). Two hundred forty-one (51.9%) participants were women, and 223 (48.1%) were men. Three hundred fifteen (68%) students were freshmen, 85 (18%) were sophomores, and 64 (14%) were juniors. The majority of the participants identified themselves as White (54.7%), with an additional 24.6% Asian, 6.9% Hispanic, 3.9% Black, and 6.0% other, and with 3.4% choosing not to respond. Of the 464 participants, 382 (82.3%) were born in the United States, and 82 (17.7%) were not. Students also provided information about their native language. Three hundred seventy-one students (80%) reported being native English speakers and 93 (20%) native speakers of a language other than English.

Instrumentation

Examination. As a part of course requirements, students were asked to write a 500-word expository essay demonstrating their understanding of theories of motivation that were part of their readings and class discussions. Their score on this essay served as a component of their overall grade in the course. Before the topic was presented, students received the following instructions:

> Now you're ready to write the essay. Below is the rubric which explains how the essay will be evaluated. The rubric will be available to you while writing. YOUR ESSAY MUST NOT EXCEED 500 WORDS. You must type your essay directly into the page. You may not cut and paste from Microsoft Word or any other software. Good luck!

The prompt for this examination was a modification of an Educational Testing Service (ETS) prompt developed for their E-Rater essay scoring system (Attali & Burstein, 2006; Burstein, 2003) deemed appropriate for first-year students that incorporated a reference to theories of motivation:

> Sometimes we choose to do things that we do not really enjoy—studying hard, eating the right foods, and so on. Describe something you do by choice that you really do not enjoy. Using theories of motivation, explain why you might continue to do it. Discuss the changes that might occur in your life if you were to stop this activity.

> Support your claims with specific examples from your life and the course reading.

Students were presented with an extensive rubric describing the criteria for evaluation. The rubric was available during the task and could be consulted at any point in the writing process. To make sure that students wrote essays of comparable length, a real-time indicator displayed a word count. The primary dependent measure used in the analyses was students' final score on the examination. Their preliminary score, prior to receiving feedback, served as a covariate in the design. A detailed description of the scoring procedures is presented below.[1]

Test motivation measure. The Post-Test Index of Test Motivation (Wolf & Smith, 1995) was used to test how motivated students were to do well on the task in question. This measure is different from other motivation measures in an important respect: it is test-specific, in that the items refer specifically to the test that has just been taken. The scale consists of eight 7-point Likert-type items bounded by (1) *strongly disagree* and (7) *strongly agree*. A sample item typical of the measure is, "Doing well on this exam was important to me." High scores on the scale indicate that students had a strong desire to do well on the exam that they just took and exerted all the necessary effort to ensure success. Lower scores suggest a lack of interest in the process or the outcome of the exam. Reliability coefficients reported in the literature are .89 (Spencer, 2005) and .87 (Wolf et al., 1995), which are similar to the $\alpha = .85$ found in the current study.

Test self-efficacy measure. The Post-Test Self-Efficacy Scale is modeled on the Post-Test Index of Test Motivation (Wolf & Smith, 1995), in that it focuses on an individual's sense of self-efficacy on a test that has just been completed. It consists of eight Likert-type items (Spencer, 2005). The answers were based on a 7-point response scale ranging from (1) *strongly disagree* to (7) *strongly agree*. A sample item typical of the measure is, "I am not competent enough to have done well on this exam" (scoring reversed). This measure assesses students' judgment of their own capabilities for the test they have completed. Higher scores on the measure indicate students' confidence in their performance on the test, and lower scores suggest doubt in their ability to have done well on the test in question. The reported alpha coefficient of the instrument is .86 (Spencer, 2005), identical to $\alpha = .86$ found in the present inquiry.

Measure of affect. The Positive and Negative Affect Scale (PANAS) is a 20-item self-report measure of positive and negative affect (Watson, Clark, & Tellegen, 1988). The scale is accompanied by instructions for measuring students' current affective state. The participants were asked to indicate the extent to which they experienced the affective states described by the PANAS adjectives on a 5-point scale ranging from (1) *slightly/not at all* to (5) *extremely*. Two additive indices were computed, resulting in separate positive affect and negative affect scores for each participant. The reported alpha coefficients of the positive affect scale range from .86 to .95, and the negative affect scale from .84 to .92 (Crawford & Henry, 2004; Ilies & Judge, 2005; Jolly, Dyck, Kramer, & Wherry, 1994; Roesch, 1998). We obtained coefficients of $\alpha = .89$ and $\alpha = .86$, respectively.

Helpfulness and accuracy of feedback. Two items were used to gauge participants' perceptions of accuracy and helpfulness of feedback: "How accurate was the feedback?" and "How helpful was the feedback?" The answers were based on a 7-point response scale ranging from (1) *not at all accurate* (*helpful*) to (7) *very accurate* (*helpful*).

Procedure

The experiment involved computer administration and was conducted in two sessions separated by 1 week. A data collection program and an interactive Web site were created to satisfy specific requirements of the study. Students were informed of the nature of the study and told that participation in the study would satisfy their psychology subject pool requirement. They were also told that all final test scores would be adjusted so that the means of all groups would equal the mean of the highest scoring group in the experiment. Thus, there would be no detriment to their grade for having participated in the study. Students were reminded that they could choose not to allow their responses to be used for research purposes. If they chose to do so, they were asked to complete the requirements of the exam and not fill out additional assessments.

First session. All students who were enrolled in the two introductory psychology courses were scheduled to come to a computer lab to take their examination. Students were presented with the test instructions and the grading rubric and were then asked to begin their essay. Students submitted their work—which was saved in the system—were thanked for their participation, and reminded to return in 1 week for the second part of the study.

Scoring of the examination. ETS allowed the use of their proprietary software package E-Rater for this study. E-Rater (Attali & Burstein, 2006) extracts linguistically based features from an essay and uses a statistical model of how these features are related to overall writing quality to assign a holistic score to the essay. Additionally, it assesses and provides feedback for errors in grammar, usage, and mechanics, identifies the essay's structure, recognizes undesirable stylistic features, and provides diagnostic annotations within each essay (Attali, 2004; Burstein, 2003). The total examination score presented to the students comprised two separate components: the E-Rater score (ranging from 0 to 6) and the content score provided by the instructor and the researcher (ranging from 0 to 6, including half points). The final score was calculated as a weighted average of the two scores and converted into a scale of 100. The E-Rater score contributed 30% to the total score, and the content score contributed 70% to the total score.

E-Rater was customized to rate the essays written on the prompt selected for the present study. Students' essays were scored on all of the aforementioned characteristics including mechanics, grammar, spelling, and stylistic features, and a holistic score was assigned to every student. For several experimental conditions, the feedback provided by E-Rater was modified to satisfy the requirements of specific feedback conditions described below. A portion of the detailed feedback screen is presented in Figure 1.

[1] Exploratory and confirmatory factor analyses of the three measures have been conducted. The results replicated previous findings reported in the literature and demonstrated the theoretical and psychometric soundness of the three measures. Because of space limitations, the results of the analyses have been excluded from this article. They are available upon request from Anastasiya A. Lipnevich.

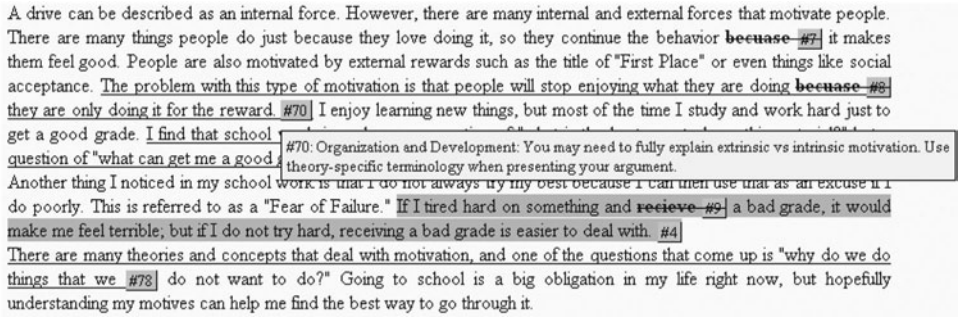

Figure 1. Detailed feedback screen with a pop-up message for a specific feedback item.

Additionally, two raters (the course instructor and the researcher) scored the content aspect of the examination. Prior to scoring the main experiment, a series of calibration sessions were held to ensure interrater reliability between the two raters. We developed a detailed rubric that provided criteria for evaluating the content of students' essays (see the Appendix). The interrater reliability was .96 for the first session examination score and .98 for the final examination score. In case of a discrepancy in ratings, the average of the two raters' scores was taken. There were no differences in ratings larger than 1 point. The instructor and the researcher were unaware of the students' identities and experimental conditions.

To provide feedback on the content of students' essays in a consistent fashion, a number of standard comments were written. These comments were slightly modified depending on the experimental condition, so that some comments sounded as if they came from a computer and others from the professor. The comments presented to each individual student reflected their particular mistakes and omissions and therefore were highly specific to each individual's work. The combination of the E-Rater essay feedback and the content feedback generated by the instructor and the researcher are referred to hereinafter as *detailed feedback*. By detailed feedback, we mean feedback that is extensive and that relates to sentence and phrase level writing as well as commentary on the quality of the content of the essay. After the initial essays were scored, blocking was used to assign participants to three experimental conditions so that the resulting groups had equivalent numbers of students with high, medium, and low first-session scores.

Each student was assigned to one of the three detailed feedback conditions:

1. No-Feedback Condition. This group received no detailed feedback.

2. Instructor-Feedback Condition. This group received a combination of the E-rater–generated feedback regarding mechanics and style and content-related comments and suggestions. Students were informed that the feedback came from the course instructor. All comments were written in a reserved and neutral fashion but in a way that was clear that they came from a person rather than a computer. To make sure that the source of feedback was clear to the participants, a clip-art picture of a typical college professor was displayed in the corner of every exam screen, and the following instructions were provided: "During this session, you will be able to edit and improve the essay you wrote the first time, based on detailed feedback I have given you on content, grammar, punctuation, spelling, sentence structure, and the overall quality of your essay. PLEASE READ MY COMMENTS CAREFULLY and do your best to use them — it should really help you get a better score."

3. Computer-Feedback Condition. Students in this group received feedback equivalent in its nature to the one in the previous condition (i.e., all of the comments were work specific and directly linked to students' essays). In this condition, students were told that all the comments were generated by the computer. The following instructions were provided: "During this session, you will be able to edit and improve the essay you wrote the first time, based on detailed feedback generated by an intelligent computer system designed to read and critique essays. The computer will give you feedback on content, grammar, punctuation, spelling, sentence structure, and the overall quality of your essay. PLEASE READ THE COMPUTER'S COMMENTS CAREFULLY and do your best to use them — it should really help you get a better score." A picture of a computer was displayed on every screen. The E-Rater comments were taken in their original form. The additional comments concerning the content and adequacy of the use of course-related constructs matched the style of the computer comments and were impersonal and neutral. A comparative table of the comments received by students in the computer and instructor conditions is presented in Table 1.

Additionally, the three feedback conditions were crossed with two factors of grade (grade or no grade) and praise (praise or no praise) resulting in a $3 \times 2 \times 2$ experimental design.

Numeric grades for the first draft of the essay were presented only to those students in the "grade" condition. Students to whom their first session score was revealed were informed that this preliminary score was only for information, and it was the final score on the revised essay that was to be counted as their outcome. Praise was provided in the form of a comment preceding the rest of the feedback. There were three levels of praise that differed depending on the score that students received for the draft of their essay (whether they were presented with this score or not). These

Table 1
Comparison of Comments Received by Students in the Instructor and Computer Conditions

Type of comment	Instructor	Computer
Mechanics	*Name*, please break your essay into paragraphs so I can see the structure.	Please break your essay into paragraphs so that the structure can be detected.
	Name, this sentence is a fragment. Proofread the sentence to be sure that it has correct punctuation and that it has an independent clause with a complete subject and predicate.	This sentence may be a fragment. Proofread the sentence to be sure that it has correct punctuation and that it has an independent clause with a complete subject and predicate.
	Name, these sentences begin with coordinating conjunctions. Try to combine the sentence that begins with *but* with the sentence that comes before it.	These sentences begin with coordinating conjunctions. A sentence that begins with *and*, *but*, and *or* can sometimes be combined with the sentence that comes before it.
Content	*Name*, a good essay usually contains three main ideas, each developed in a paragraph. Use examples, explanations, and details to support and extend your main ideas. Try to center them around the theories of motivation I discussed in class. Include details and theory-specific terminology.	A good essay usually contains three main ideas, each developed in a paragraph. Use examples, explanations, and details to support and extend your main ideas. Center them around the theories of motivation. Include details and theory-specific terminology.
	Name, please discuss all of the components of the Drive reduction theory: need, drive, action, and homeostasis. You are missing two of the components.	You may need to discuss all of the components of the Drive reduction theory: need, drive, action, and homeostasis.
	Name, discuss all of the components of Atkinson's theory: expectancy, value and the need for achievement. You are missing one of the components.	Discuss all of the components of Atkinson's theory: expectancy, value and the need for achievement. You may be missing some of the components.

levels were used to avoid having students receive a praise statement clearly incongruous to their level of performance. See Table 2 for the three levels of praise used.

Second Session

Participants were asked to return to the computer lab 1 week after taking the initial examination. They logged into the system and were shown their essays with corresponding feedback. What appeared to students on the computer screen (detailed feedback, praise, etc.) depended on the condition to which they had been randomly assigned. After viewing their combination of detailed feedback, praise, and grade (or lack thereof), but prior to moving to the essay revision screen, students were asked to fill out the Positive Affect Scale and the Negative Affect Scale. The participants were then prompted to make revisions and resubmit their essay on the basis of the feedback they received. Students could refer to the grading rubric and to their feedback comments at any point of the session by hovering the mouse over hotspots in the feedback text. A portion of the detailed feedback screen is presented in Figure 1.

Students who did not receive detailed feedback, praise, or grades were encouraged to reread their essays, consult the rubric, and work on improving their work. After participants submitted their revised essays, they were asked to make a judgment concerning the accuracy and helpfulness of the feedback. They were then asked to complete the Post-Test Index of Test Motivation, and the Post-Test Self-Efficacy Scale.

Scoring of the revised essay followed the rules of the first draft scoring. The final numeric grade was computed as a weighted mean of the E-Rater (30%) and the content (70%) score. The

Table 2
Levels of Praise for the Instructor, Computer and No-Feedback Conditions

Exam score	Instructor feedback	Computer feedback	No feedback
80 to 100	*Name*, you made an excellent start with this essay! I still see room for improvement, so take some time and make it even better.	You made an excellent start with this essay. The data indicate there is still room for improvement, so take some time and make it even better.	You made an excellent start with this essay! There is still room for improvement, so take some time and make it even better.
70 to 79	*Name*, you made a very good start with this essay! I still see room for improvement, so take some time and make it better.	You made a very good start with this essay. The data indicate there is still room for improvement, so take some time and make it better.	You made a very good start with this essay! There is still room for improvement, so take some time and make it better.
69 and below	*Name*, you made a good start with this essay! I still see room for improvement, so take some time and make it better.	You made a good start with this essay. The data indicate there is still room for improvement, so take some time and make it better.	You made a good start with this essay! There is still room for improvement, so take some time and make it better.

scorers were blind to unaware of student identity and experimental condition.

After the completion of the study, a series of focus groups was held with 50 students to explore their reactions to the experiment and to ensure that all students understood the nature of the conditions they were in. The results indicated that all participants understood that they were either getting feedback from a computer program or from their course instructor. A schematic representation of the steps in the study is presented in Figure 2.

Results

Means, standard deviations, and intercorrelations of all major variables in the study are presented for purposes of reference in Table 3.

Analyses of the Effects of Treatments on the Final Exam Score

A 3 × 2 × 2 analysis of covariance (ANCOVA) with detailed feedback (3 levels), grade (2 levels), and praise (2 levels) conditions as factors and the grade for the first essay draft (before revisions) as a covariate, examined differences in the final numeric grades for the essay exam. A Bonferroni adjustment was used to control for Type I error (the criterion used was .0083). Significant main effects were found for detailed feedback and for grade but not for praise. Also, there were significant interaction effects found for grade and praise, as well as for grade and detailed feedback. No other interactions were significant. The effect of detailed feedback was strong; the effect of grade was moderate and needs to be examined in light of the two small, but significant, interactions involving grade. We examine the main effect of detailed feedback first and then the intriguing combination of effects involving presentation of grades.

There was a strong significant main effect of detailed feedback on students' final score, $F(2, 450) = 69.23$, $p < .001$, $\eta^2 = .24$. Post hoc analyses show that students who did not receive detailed feedback obtained substantially lower final exam scores than did those who received detailed feedback from either the computer ($p < .01$) or the instructor ($p < .01$), and there were no differences in students' performance between computer and instructor conditions ($p > .05$; see Table 4 for means). Differences between the no-detailed-feedback condition and the two detailed-feedback conditions showed effect sizes (Cohen's *d*) of between .30 to 1.25, depending on the presence of grade and praise.

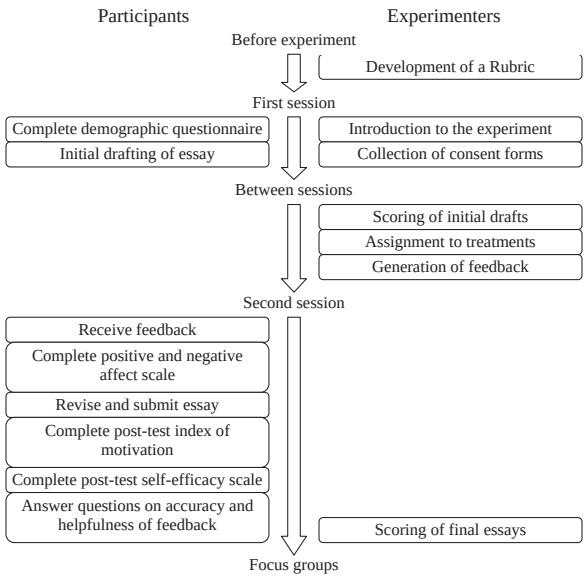

Figure 2. Schematic of procedures and administration of measures.

There was also a significant difference in the final exam score between students in the grade condition and those in the no-grade condition, $F(1, 450) = 4.07$, $p < .05$, $\eta^2 = .04$. Students who were shown the grade they received for their first draft performed less well on the revision than did those who were not shown their grade. This effect needs to be viewed, however, in the context of two significant interaction terms involving grade.

The analysis revealed a significant disordinal interaction between grade and praise, $F(1, 450) = 6.00$, $p < .05$, $\eta^2 = .04$. Students in the no-grade/no-praise condition received the highest scores ($M = 79.82$, $SD = 5.12$). The lowest scores were observed in the grade/no-praise condition ($M = 77.69$, $SD = 5.12$). The grade/ praise condition also produced fairly high scores ($M = 79.26$, $SD = 5.12$), as did the no-grade/praise condition ($M = 79.06$, $SD = 5.13$). Means and standard deviations are presented in Table 5. Although the cell means cannot be directly compared given that they are interaction terms, the simplest explanation of this interaction appears to be that the presentation of grades depressed performance unless ameliorated by the presence of a statement of praise.

There was also a significant interaction between grade and detailed feedback, $F(2, 450) = 5.54$, $p < .01$, $\eta^2 = .08$. In the no-detailed-feedback condition, scores were fairly similar for students who received a grade ($M = 75.37$, $SD = 5.12$) in comparison with those who did not receive a grade ($M = 74.65$, $SD = 5.12$). Under the computer detailed-feedback condition, students' scores were again similar ($M = 80.44$, $SD = 5.12$, for the no-grade condition, to $M = 80.93$, $SD = 5.12$, for the grade condition), but under the instructor detailed-feedback condition, a distinct difference was observed. Students' final exam scores were relatively high when their grade was not presented ($M = 82.74$, $SD = 5.13$) and was substantially lower for students when their grade was presented ($M = 79.63$, $SD = 5.12$). Means and standard deviations are presented in Table 6.

In summary, the analysis of the performance scores supported the first hypothesis about the overall positive effect of detailed feedback on students' improvement. There were no differences for perceived source of the feedback. Therefore, the hypothesis about the differential effect of computer feedback was not supported. Receipt of a numeric grade led to a substantial decline in performance, especially for students who thought the grade had come from the instructor. However, a praise statement appeared to lessen that effect. The hypothesis positing that presentation of a grade hinders improvement was supported, whereas the hypothesis about negative effect of praise was not, because no main effect of praise was found. But presenting praise appeared to lessen the negative effect of presenting a grade.

Analysis of Differences in the Final Exam Score by Students' Initial Performance

Following Butler (1988), we decided to investigate whether the differences found for the overall analysis would be replicated if we

Table 3
Descriptive Statistics and Intercorrelations of Study Variables

Variable	*M*	*SD*	1	2	3	4	5	6
1. Draft grade	74.42	8.28	—					
2. Final grade	78.94	8.72	.74***	—				
3. Positive Affect Scale	29.86	7.17	.02	−.02	—			
4. Negative Affect Scale	24.00	7.51	−.14**	−.06	−.06	—		
5. Posttest Index of Test Motivation	48.19	6.79	.09*	.11*	.30***	.08	—	
6. Test Self-Efficacy Scale	44.44	6.77	.24***	.23***	.29***	−.22***	.37***	—

Note. For the Self-Efficacy and Positive Affect Scales, $N = 462$. For the remaining measures, $N = 463$.
* $p < .05$. ** $p < .01$. *** $p < .001$.

examined students at varying levels of performance on the initial drafts. To that end, a frequency analysis was run for the initial draft score. The analysis revealed a mean of 74.42, $SD = 8.28$, and a range from 50 to 96. The analysis of frequency tables showed that 25% of the sample scored at or below 69 (equivalent to letter grades D and F), about 50% received a score between 70 and 79 (equivalent to the letter grade C), and the remaining 25% obtained a score at or above 80 (equivalent to letter grades B and A). On the basis of these cut points, students were identified as having low ($N = 116$), medium ($N = 217$), and high ($N = 130$) initial draft scores. We split the dataset on the first exam score grouping variable and ran a series of $3 \times 2 \times 2$ ANCOVAs with the detailed feedback ($\times 3$), grade ($\times 2$), and praise ($\times 2$) as factors and with the first session grade as a covariate. These analyses examined differences in the final exam scores for students in each initial performance group.

Students With Low Initial Draft Score

For students who received low scores on their initial draft, the analysis revealed a significant Grade $\times$Detailed Feedback interaction, $F(2, 103) = 5.27$, $p < .01$, $\eta^2 = .10$. In the no-detailed-feedback condition, scores were higher for students' who received a grade ($M = 67.85$, $SD = 6.64$) than for those who did not receive a grade ($M = 64.15$, $SD = 6.75$). The overall scores were quite low for these groups. Under the computer detailed-feedback condition, students' scores were higher when the grade was presented ($M = 75.50$, $SD = 6.71$) than when no grade was presented ($M = 72.07$, $SD = 6.64$). Under the instructor detailed-feedback condition, students' final exam scores were relatively high for the no-grade condition, but they were lower when the grade was presented ($M = 77.24$, $SD = 6.86$, when no grade was presented; $M = 72.07$, $SD = 6.65$, when grade was presented). See Table 7 for means and standard deviations.

There was also a significant effect for the detailed feedback, $F(2, 103) = 18.78$, $p < .001$, $\eta^2 = .28$, with students in the control condition (who received no detailed feedback) scoring significantly lower ($M = 65.46$; $SD = 6.06$) than those in either the instructor ($M = 75.11$, $SD = 6.94$, $p < .01$) or computer conditions ($M = 73.88$, $SD = 8.04$, $p < .01$). No differences were revealed between the computer and instructor detailed-feedback conditions ($p > .05$), and no significant effects were found for grade, praise, or for the other interaction terms.

Students With Medium Initial Draft Score

For students who received a medium draft score (between 70 and 79), a significant effect for the detailed feedback, $F(2, 204) =$

Table 4
Means and Standard Deviations of the Final Exam Scores by Detailed Feedback, Grade, and Praise

	No grade			Grade			Total		
Condition	No praise	Praise	Total	No praise	Praise	Total	No praise	Praise	Total
No feedback									
M	73.80	74.38	74.09	75.11	76.24	75.67	74.44	75.27	74.85
SD	8.57	9.21	8.84	8.56	7.60	8.07	8.54	8.47	8.49
N	40	40	80	38	37	75	78	77	155
Computer									
M	81.15	79.75	80.44	79.80	80.28	80.04	80.47	80.01	80.24
SD	8.43	8.97	8.68	7.07	8.36	7.70	7.75	8.62	8.18
N	39	40	79	40	40	80	79	80	159
Instructor									
M	83.85	83.26	83.57	78.41	81.74	80.09	81.20	82.47	81.82
SD	7.60	7.56	7.53	7.84	7.92	8.01	8.14	7.74	7.94
N	39	35	74	37	38	75	76	73	149
Total									
M	79.55	78.95	79.25	76.80	79.16	78.63	78.69	79.20	78.94
SD	9.20	9.32	9.24	8.02	8.24	8.15	8.66	8.78	8.71
N	118	115	233	115	115	230	233	230	463

Table 5
Estimated Marginal Means and Standard Deviations of the Final Exam Score by Grade and Praise

Condition	*M*	*SD*	*N*
No grade			
No praise	79.82	5.12	118
Praise	79.06	5.13	115
Grade			
No praise	77.69	5.12	115
Praise	79.26	5.12	115

Note. Adjusted means after controlling for the first exam score.

34.87, $p < .001$, $\eta^2 = .26$, was found. Pairwise comparisons revealed that students in the control condition scored significantly lower ($M = 74.23$, $SD = 4.79$) than did those in either instructor condition ($M = 80.23$, $SD = 6.33$, $p < .01$) or computer condition ($M = 79.54$, $SD = 5.29$, $p < .01$). No differences were found between the instructor and the computer conditions ($p > .05$). Additionally, significant differences were found between participants in the grade and no-grade conditions, $F(1, 204) = 7.9$, $p < .001$, $\eta^2 = .09$. Students who were shown their initial draft grade scored lower than did those who were not shown their grade ($M = 76.06$, $SD = 5.54$, for the grade condition; $M = 78.88$, $SD = 6.03$, for the no-grade condition). Grade ×Detailed Feedback was found not to be significant for this group of students.

Students With High Initial Draft Score

For the high-scoring group (80 and above), the ANCOVA revealed a significant effect for the detailed feedback, $F(2, 117) = 18.13$, $p < .001$, $\eta^2 = .24$, with students in the control condition scoring significantly lower ($M = 84.49$, $SD = 4.88$) than did those in either the instructor condition ($M = 88.49$, $SD = 5.14$, $p < .01$) or computer condition ($M = 88.76$, $SD = 4.35$, $p < .01$). No differences were found between the computer and instructor detailed-feedback conditions ($p > .05$). Additionally, significant differences were found between the grade and no-grade conditions, $F(1, 117) = 3.72$, $p < .01$, $\eta^2 = .05$. High-scoring students in the grade condition scored significantly lower than did those in the no-grade condition ($M = 86.54$, $SD = 4.95$, for the grade condition; $M = 88.25$, $SD = 5.18$, for the no- grade condition).

Table 6
Estimated Marginal Means and Standard Deviations of the Final Exam Score by Grade and Detailed Feedback

Condition	*M*	*SD*	*N*
No grade			
No feedback	74.65	5.12	80
Computer	80.93	5.12	79
Instructor	82.74	5.13	74
Grade			
No feedback	75.37	5.12	75
Computer	80.43	5.12	80
Instructor	79.63	5.12	75

Note. Adjusted means after controlling for the first exam score.

Table 7
Estimated Marginal Means and Standard Deviations of the Final Exam Score by Grade and Source of Feedback for Low-Ability Students

Condition	*M*	*SD*	*N*
No grade			
No feedback	64.15	6.75	19
Computer	72.07	6.64	21
Instructor	77.24	6.86	18
Grade			
No feedback	67.85	6.64	18
Computer	75.50	6.71	21
Instructor	72.07	6.65	19

Note. Adjusted means after controlling for the first exam score.

Overall, the analyses showed that students who scored low on the first draft responded favorably to detailed feedback and were able to improve upon it. However, when presented with a grade from the instructor, these students did not do as well as when they were oblivious to their draft grade. At the same time, we found that low-scoring students did not react negatively to a grade if they believed it had come from the computer or when a grade was the only feedback they received. Both medium and high scorers were shown to respond well to detailed feedback coming from either the computer or the instructor. Their performance, however, depended on whether a grade was presented, with those who received a grade scoring lower than did those who did not. It did not matter whether the grade came from the computer or the instructor, as students' response to it was comparably unfavorable.

Analyses of Differences in Motivation, Self-Efficacy, and Affect

The relationships among detailed feedback, praise, and grades, and students' motivation, self-efficacy, and negative and positive affect were investigated via two $3 \times 2 \times 2$ multivariate analyses of variance (MANOVAs). The first MANOVA included self-efficacy and motivation as dependent variables, and grade, praise, and detailed feedback as independent variables. The second MANOVA was run with Positive Affect Scale and Negative Affect Scale scores as dependent variables, and with grade, praise, and detailed feedback as independent variables. We ran the two analyses separately because the data for them were gathered at different points in the experiment.

For self-efficacy and motivation, multivariate tests were significant for the grade factor—the F statistic for Wilks' lambda was $F(2, 449) = 5.42$, $p < .01$—and for the praise factor—the F statistic for Wilks' lambda was $F(2, 449) = 4.02$, $p < .01$—but not for the detailed feedback or any of the interactions. To test the difference for both of the dependent variables, univariate analyses were performed for motivation and self-efficacy.

For motivation, the univariate results indicate significant differences in motivation levels between students who received praise on their initial performance and those who did not, $F(1, 450) = 7.58$, $p < .01$, $\eta^2 = .04$. Interestingly, students in the praise condition reported lower motivation ($M = 47.29$, $SD = 7.66$) than did students in the no-praise condition ($M = 49.06$, $SD = 5.71$).

For self-efficacy, the results indicated a significant grade effect, $F(1, 450) = 10.80$, $p < .01$, $\eta^2 = .08$, with students who received a grade for their initial draft exhibiting lower self-efficacy levels ($M = 43.38$, $SD = 7.03$) than did those who were unaware of their draft grade ($M = 45.47$, $SD = 6.36$).

For positive and negative affect, multivariate tests were only significant for the grade factor. The F statistic for Wilks' lambda was $F(2, 450) = 7.03$, $p < .01$. To test the difference for both of the dependent variables, univariate analyses were performed for both positive and negative affect variables. Similarly to self-efficacy, there was a significant difference in negative affect depending on the presence or absence of grade, $F(1, 450) = 14.09$, $p < .01$, $\eta^2 = .08$. Students who received a grade for their draft reported higher levels of negative affect ($M = 25.27$, $SD = 7.68$) than did those who did not receive their grade ($M = 22.72$, $SD = 7.12$). For positive affect, there were no significant effects for any of the independent variables or their interactions.

Overall, presence of grade was shown to have a significant effect on students' reported self-efficacy and negative affect. Students who received a grade had more negative affect and reported lower levels of self-efficacy than did their counterparts for whom their grade was unknown. Praise affected motivation, but in an unexpected fashion, with students who were presented with a laudatory statement reporting lower levels of motivation than did those who were not.

Analyses of Differences in Perceived Helpfulness and Accuracy of Feedback

To examine the final issue addressed in the study about the perceived helpfulness of detailed feedback and perceived accuracy of that feedback, a $3 \times 2 \times 2$ MANOVA was used. Perceived helpfulness and accuracy of detailed feedback were used as dependent variables, and grade, praise, and the detailed feedback as independent variables. Multivariate analyses only revealed significant effects for the detailed feedback; the F statistic for Wilks' lambda was $F(4, 900) = 87.10$, $p < .001$.

Subsequent univariate analyses with the perceived accuracy of detailed feedback as dependent variable revealed a significant effect for the detailed feedback factor, $F(2, 451) = 130.98$, $p < .001$, $\eta^2 = .37$. A post hoc analysis yielded a significant difference in accuracy ratings between instructor and computer conditions ($p < .01$), between instructor and no-detailed feedback conditions ($p < .01$), and between the computer and no-feedback conditions ($p < .01$). Students who received their feedback that was perceived to come from the instructor rated feedback as being more accurate ($M = 5.95$, $SD = 1.07$) than did those who received feedback perceived to be from a computer ($M = 5.33$, $SD = 1.42$) or those who did not receive detailed feedback ($M = 3.30$, $SD = 1.91$). Of course, those receiving no detailed feedback had little basis for making a judgment.

Univariate analysis with perceived helpfulness of feedback revealed a significant effect for the detailed feedback, $F(2, 451) = 206.12$, $p < .001$, $\eta^2 = .48$. A post hoc analysis ($p < .01$) indicated a significant difference in helpfulness of feedback ratings between the instructor and computer conditions, between the instructor and no-feedback conditions, and between the computer and no-feedback conditions. Students who received feedback from the instructor rated it as being more helpful ($M = 6.06$, $SD = 1.07$) in comparison with those students who believed that feedback was computer generated ($M = 5.44$, $SD = 1.56$) or those who did not receive detailed feedback ($M = 2.79$, $SD = 1.76$).

Overall, students rated detailed feedback from the instructor as more helpful and accurate than did students in the computer feedback condition. The presence of grade or praise did not affect the perceptions of the accuracy or helpfulness of the feedback.

Discussion

The strongest and most consistent finding of the study was that written, detailed feedback specific to individual work was strongly related to improvement. The effects of grades and praise on performance were more complex. Students in the instructor feedback group who also received a grade for their draft had lower scores than did those who did not receive a grade. However, if they received a grade and a statement of praise, the negative effect was ameliorated. It is interesting to note that the highest-performing group in the study was the one receiving detailed feedback perceived to come from the instructor with no grade and no praise.

These findings are consistent with the research showing that descriptive feedback that conveys information on how one performs the task and details ways to overcome difficulties is far more effective than is evaluative feedback, which simply informs students about how well they did (Hattie & Timperley, 2007; Kluger & DeNisi, 1998). Indeed, across the entire sample, for students of all writing ability levels, detailed feedback led to the greatest improvement. The importance of detailed feedback is especially clear for tasks that are loosely framed and do not have a simple right or wrong answer (Bangert-Drowns et al., 1991; Roos & Hamilton, 2005).

Differences in Responses Depending on the Perceived Source of Feedback

We found no significant differences due to source of feedback. This finding provides partial support for the "computers as social actors" paradigm, suggesting that people may be unconsciously perceiving computers as "intentional social agents," and because of this, computer-provided feedback tends to elicit the same or very similar responses from individuals (Mishra, 2006; Nass et al., 1996, 1999). The support for this paradigm is only partial, because although students' exam scores were quite similar for both computer and instructor conditions, interactions between the source of feedback and grade and praise were consistently found.

The competing paradigm, which proposes that computers are generally perceived as neutral tools (Earley, 1988; Lepper et al., 1993), was not supported here. According to this perspective, computers tend to be viewed as neutral and unbiased sources of information, and feedback received from computers is more trusted by individuals. Quite contrary to this viewpoint, participants in our study rated the instructor's feedback as being more accurate and helpful than was computer-generated feedback.

Effects of Grades on Student Performance

The effect of receiving a grade in this study was particularly interesting. There was a main effect for grade and two notable interactions. Among those students who believed that they re-

ceived their detailed feedback from the instructor, those who were given a grade for their draft showed substantially lower scores than did those who were not. Receiving a grade was also generally associated with lower self-efficacy and more negative affect. One explanation for theses findings comes from the feedback intervention theory of Kluger and DeNisi (1996). They suggested that optimal feedback should direct individuals' attention toward the task and toward the specific strategies that would lead to achievement of desired outcomes. Letter grades or numeric scores, being evaluative in nature, tend to turn students' attention away from the task and toward the self, leading to negative effects on performance (Kluger & DeNisi, 1996; Siero & Van Oudenhoven, 1995; Szalma, Hancock, Warm, Dember, & Parsons, 2006). The findings are also consistent with Hattie and Timperley's (2007) argument that feedback focused on the task and not the individual is more effective.

Similarly, attention to the self, elicited by the presentation of a grade, could activate affective reactions. According to Kluger, Lewinsohn, and Aiello (1994), feedback gets cognitively evaluated with respect to two dimensions: harm versus benefit and the need to take action. The appraisal of harm or benefit potential for the self is reflected in the primary dimension of mood (pleasantness), whereas the need to take action is reflected in a secondary dimension of mood (arousal). The affective measure administered in this study addressed the arousal dimension of mood. High positive affect was indicative of high arousal, and high negative affect was indicative of depression and behavior inhibition (Crawford & Henry, 2004). The results indicated that students who were shown their draft grade scored significantly higher on the Negative Affect Scale than did their counterparts who did not receive their draft grade. Thus, the effect of the grade may have led students to become depressed about their performance, leading them to be less disposed to put forth the necessary effort to improve their work. This effect may have been particularly strong if the grade was perceived to be coming from the instructor (as opposed to being computer generated), hence the large negative impact of grade on performance in that condition.

The negative effect of grades on students' performance can also be explained through their influences on students' self-efficacy. Self-efficacy has been shown to be influenced by prior outcomes (Bandura & Locke, 2003). Feedback, therefore, has a potential of affecting self-efficacy. The current study revealed that presentation of grade resulted in decreased levels of self-efficacy with regard to the exam. Students who were not shown their draft grade reported higher levels of exam-specific self-efficacy than did those to whom a grade was provided.

Effects of Praise on Student Performance

Our study attempted to clarify the effect of praise on students' performance, motivation, self-efficacy, and affect. Praise is a controversial topic, with some researchers arguing that praise promotes learning by raising positive affect and self-efficacy (Alber & Heward, 2000), whereas others stipulate that it leads to depletion of cognitive resources by taking attention away from the task and focusing it on aspects of the self (Baumeister et al., 1990; Kluger & DeNisi, 1996). This study did not reveal any consistent overall differences in performance among students who did or did not receive praise on their performance. Comments and grades had a stronger influence on students' performance, with praise adding to and modifying their effects. Specifically, we found that praise mitigated the adverse effect of grades on students' performance.

The only outcome measure directly affected by praise was motivation. The effect of praise here was quite interesting, if not surprising. Students presented with praise reported slightly lower levels of motivation than did their counterparts who were not praised on their initial performance (effect size of .27). The only study that somewhat agrees with the finding here was conducted by Butler (1987). The researcher demonstrated that students receiving praise on their performance reported high levels of ego involvement, decreased levels of task involvement, and higher perceptions of success while exhibiting modest performance on a task in comparison with students who were not praised on their work.

The motivation measure administered in our study did not gauge different types of motivation. It is possible that this general motivation measure corresponded to the task-involvement measure used by Butler (1987) and therefore elicited similar responses. Students presented with praise were not as interested in the task and were not as motivated to try harder, believing perhaps that they had achieved enough. This supposition could be confirmed if students' performance reflected it; however, praise appears to have a less direct—rather, a mitigating—effect on students' performance. Further research is needed.

Difference in Responses to Feedback as Dependent on Students' Draft Score

Several researchers propose that students' responses to feedback messages may depend on their ability or typical performance levels (Black & Wiliam, 1998). Very few studies have examined the differential effects of feedback on students' performance for students of different past performance. In the present study, low-, medium-, and high-scoring students on the initial essay draft showed a significant increase in scores when presented with detailed feedback. It did not matter what level their original performance was; students who were offered feedback specific to their own work found ways to incorporate it into their essay and improve their results. After covariate adjustment for pretest performance, feedback accounted for 28% of variance in the final exam score for students within the low-achievement group, and for 26% and 24% of those in the medium and high groups, respectively. Thus, the positive effect of personalized feedback was observed irrespective of students' initial writing scores.

Although detailed feedback was conducive to learning in students of all performance levels, some differences in students' responses to feedback were found between the low-scoring group on one hand and the medium- and high-scoring groups on the other. Butler (1988) showed that presentation of a grade on its own or in combination with any other information leads to a significant decline of interest in performing the task for low-achieving students. In the current study, students who received high or medium initial scores performed less well on the revision than did students in the no-grade condition. As was suggested in preceding sections, a grade appeared to undermine the effort that students were willing to put forward to improve their work. However, no overall differences between the grade and no-grade conditions were found for the low-scoring students. Instead, there was a strong Grade $\times$ De-

tailed Feedback interaction. Specifically, students receiving grades for their draft performed better in the no-detailed-feedback and computer feedback conditions but worse in the instructor feedback condition. It may be the case that the computer-based grade was viewed as being less judgmental or personally directed than was the instructor-based grade.

Limitations

Some limitations of the study should be noted. One of the feedback conditions in the study involved presentation of praise. The decision was made to use standard laudatory comments differentiated according to three levels of the quality of initial students' work. No main effects were found for the praise factor. It is possible that none of the levels of praise were strong enough to induce the responses that are commonly reported in the literature (Baumeister et al., 1990; Delin & Baumeister, 1994; Henderlong & Lepper, 2002). Comments that were more detailed and personal might have induced more positive responses from the participants. At the same time, interaction effects were found between praise and grade, as well as between praise and feedback source, which indicate that the praise manipulation was successful at least to a degree.

The effects of the various conditions examined in this study may well not operate in the same fashion for all individuals. In this study, we did not directly address individual differences among participants with the exception of considering the influence of initial scores. A systematic investigation into how individuals differ with regard to response to feedback would be one particularly fruitful area for further investigation.

The sample of the present study comprised college students who were relatively uniform in their age, with the majority of the participants being first-year students. Generalizing the results of the study to wider populations should be approached with caution. Conversely, the fact that the main experimental task was a part of a normal learning experience, and was approached by participants seriously as a regular course exam, contributed to the robustness of the findings.

Finally, the experimental task involved students working on an essay and then coming back a week later to revise their work on the basis of the feedback provided at that time. In other words, the feedback was used to monitor and improve performance on an assignment carried out over a relatively brief period. The students were not assessed later, and they were not given a similar task at a later time. Therefore, the present study does not allow for inferences concerning the long-term effect of feedback on students' performance.

Conclusions and Directions for Future Research

The findings of this study show that detailed, specific, descriptive feedback that focuses students' attention on their work, rather than on the self, is the most advantageous approach to formative feedback. The benefits of such feedback occur at all levels of performance. Evaluative feedback in the form of grades may be helpful if no other options are available and can beneficially be accompanied by some form of encouragement. At the same time, grades were shown to decrease the effect of detailed feedback. It appears that this occurs because a grade reduces a sense of self-efficacy and elicits negative affect around the assessment task.

Although the present study was strengthened by conducting the research in an actual university course, we do not know whether students receiving detailed feedback on the task at hand would perform better in a subsequent task or whether presentation of a grade led to less learning or simply to less effort on the revision of the work. One clear venue for future research would be to study how differential feedback influences subsequent learning in a course. It is, of course, difficult to conduct research that would vary the nature of the feedback that students receive on a randomized basis throughout an entire course, both for practical and ethical reasons. Yet, unless we find ways to conduct rigorous research into these issues, and their many elaborations and permutations, we will not learn the most effective approaches to providing feedback and utilizing formative assessment.

References

Airasian, P. W. (1994). *Classroom assessment.* New York, NY: McGraw-Hill.

Alber, S. R., & Heward, W. L. (2000). Teaching students to recruit positive attention: A review and recommendations. *Journal of Behavioral Education, 10,* 177–204.

Attali, Y. (2004). *Exploring the feedback and revision features of the criterion service.* Paper presented at the National Council on Measurement in Education annual meeting, San Diego, CA.

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater v. 2. *Journal of Technology, Learning, and Assessment, 4,* 123–212.

Bandura, A., & Locke, E. A. (2003). Negative self-efficacy and goal effects revisited. *Journal of Applied Psychology, 88,* 87–99.

Bangert-Drowns, R. L., Kulik, J. A., & Morgan, M. T. (1991). The instructional effect of feedback in test-like events. *Review of Educational Research, 61,* 213–238.

Baumeister, R. F., Hutton, D. G., & Cairns, K. J. (1990). Negative effects of praise on skilled performance. *Basic and Applied Social Psychology, 11,* 131–148.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5,* 7–68.

Boekaerts, M., Maes, S., & Karoly, P. (2005). Self-regulation across domains of applied psychology: Is there an emerging consensus? *Applied Psychology: An International Review, 54,* 149–154.

Burstein, J. (2003). The e-rater scoring engine: Automated essay scoring with natural language processing. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 107–116). Hillsdale, NJ: Erlbaum.

Butler, R. (1987). Task-involving and ego-involving properties of evaluation: Effects of different feedback conditions on motivational perceptions, interest, and performance. *Journal of Educational Psychology, 79,* 474–482.

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

Butler, R., & Nisan, M. (1986). Effects of no-feedback, task-related comments and grades on intrinsic motivation and performance. *Journal of Educational Psychology, 78,* 210–216.

Cameron, J., & Pierce, D. P. (1994). Reinforcement, reward, and intrinsic motivation: A meta-analysis. *Review of Educational Research, 64,* 363–423.

Crawford, J. R., & Henry, J. D. (2004). The Positive and Negative Affect Schedule (PANAS): Construct validity, measurement properties and normative data in a large non-clinical sample. *British Journal of Clinical Psychology, 43,* 245–265.

Deci, E. L., Koestner, R., & Ryan, R. M. (1999). A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin, 125,* 627–668.

Delin, C. R., & Baumeister, R. F. (1994). Praise: More than just social reinforcement. *Journal for the Theory of Social Behaviour, 24,* 219–241.

Earley, P. C. (1988). Computer-generated performance feedback in the subscription-processing industry. *Organizational Behavior and Human Decision Processes, 41,* 50–64.

Ferdig, R. E., & Mishra, P. (2004). Emotional responses to computers: Experiences in unfairness, anger and spite. *Journal of Educational Multimedia and Hypertext, 13,* 143–161.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77,* 81–113.

Henderlong, J., & Lepper, M. R. (2002). The effects of praise on children's intrinsic motivation: A review and synthesis. *Psychological Bulletin, 128,* 774–795.

Ilgen, D. R., & Davis, C. A. (2000). Bearing bad news: Reactions to negative performance feedback. *Applied Psychology: An International Review, 49,* 550–565.

Ilies, R., & Judge, T. A. (2005). Goal regulation across time: The effects of feedback and affect. *Journal of Applied Psychology, 90,* 453–467.

Jolly, J. B., Dyck, M. J., Kramer, T. A., & Wherry, J. N. (1994). Integration of positive and negative affectivity and cognitive content specificity: Improved discrimination of anxious and depressive symptoms. *Journal of Abnormal Psychology, 103,* 544–552.

Kanouse, D. E., Gumpert, P., & Canavan-Gumpert, D. (1981). The semantics of praise. In J. H. Harvey, W. Ickes, & R. F. Kidd (Eds.), *New directions in attribution research* (Vol. 3, pp. 97–115). Hillsdale, NJ: Erlbaum.

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: Historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119,* 254–284.

Kluger, A. N., Lewinsohn, S., & Aiello, J. (1994). The influence of feedback on mood: Linear effects on pleasantness and curvilinear effects on arousal. *Organizational Behavior and Human Decision Processes, 60,* 276–299.

Landauer, T. K., Latham, D., & Foltz, P. (2003). Automatic essay assessment. *Assessment in Education, 10,* 124–135.

Lepper, M. R., Henderlong, J., & Gingras, I. (1999). Understanding the effects of extrinsic rewards on intrinsic motivation: Uses and abuses of meta-analysis: Comment on Deci, Koestner, and Ryan (1999). *Psychological Bulletin, 125,* 669–676.

Lepper, M. R., Woolverton, M., Mumme, D. L., & Gurtner, J. (1993). Motivational techniques of expert human tutors: Lessons for the design of computer-based tutors. In S. P. Lajoie & S. J. Derry (Eds.), *Computers as cognitive tools* (pp. 75–106). Hillsdale, NJ: Erlbaum.

Marzano, R. (2000). *Transforming classroom grading.* Alexandria, VA: Association for Supervision and Curriculum and Development.

Mishra, P. (2006). Affective feedback from computers and its effect on perceived ability and affect: A test of the computers as social actors hypothesis. *Journal of Educational Multimedia and Hypermedia, 15,* 107–131.

Nass, C., Fogg, B. J., & Moon, Y. (1996). Can computers be teammates? *International Journal of Human–Computer Studies, 45,* 669–678.

Nass, C., Moon, Y., & Carney, P. (1999). Are respondents polite to computers? Social desirability and direct responses to computers. *Journal of Applied Social Psychology, 29,* 1093–1110.

Nass, C., Moon, Y., & Green, N. (1997). Are computers gender-neutral? Gender stereotypic responses to computers. *Journal of Applied Social Psychology, 27,* 864–876.

Oosterhof, A. (2001). *Classroom applications of educational measurement.* Upper Saddle River, NJ: Merrill Prentice Hall.

Orrell, J. (2006). Feedback on learning achievement: Rhetoric and reality. *Teaching in Higher Education, 11,* 441–456.

Ramaprasad, A. (1983). On the definition of feedback. *Behavioral Science, 28,* 4–13.

Roesch, S. C. (1998). The factorial validity of trait positive affect scores: Confirmatory factor analyses of unidimensional and multidimensional models. *Educational and Psychological Measurement, 58,* 451–466.

Roos, B., & Hamilton, D. (2005). Formative assessment: A cybernetic viewpoint. *Assessment in Education, 12,* 7–20.

Shute, V. J. (2007). *Focus on formative feedback* (Rep. No. RR-07–11). Princeton, NJ: Educational Testing Service.

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

Siero, F., & Van Oudenhoven, J. P. (1995). The effects of contingent feedback on perceived control and performance. *European Journal of Psychology of Education, 10,* 13–24.

Smith, E., & Gorard, S. (2005). They don't give us our marks: The role of formative feedback in student progress. *Assessment in Education Principles Policy & Practice, 12,* 21–38.

Spencer, S. (2005). *Stereotype threat in mathematics in undergraduate women.* Unpublished doctoral dissertation, Rutgers University, New Brunswick, NJ.

Symonds, K. W. (2004). *After the test: Closing the achievement gap with data.* Naperville, IL: Learning Point Associates.

Szalma, J. L., Hancock, P. A., Warm, J. S., Dember, W. N., & Parsons, K. S. (2006). Training for vigilance: Using predictive power to evaluate feedback effectiveness. *Human Factors, 48,* 682–692.

Vancouver, J. B., More, K. M., & Yoder, R. J. (2008). Self-efficacy and resource allocation: Support for a nonmonotonic discontinuous model. *Journal of Applied Psychology, 93,* 35–47.

Vancouver, J. B., Thompson, C. M., & Williams, A. A. (2001). The changing signs in the relationships between self-efficacy, personal goals, and performance. *Journal of Applied Psychology, 86,* 605–620.

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS Scales. *Journal of Personality and Social Psychology, 47,* 1063–1070.

Wiliam, D., & Thompson, M. (2007). Integrating assessment with instruction: What will it take to make it work? In C. A. Dwyer (Ed.), *The future of assessment: Shaping teaching and learning.* Mahwah, NJ: Erlbaum.

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment, 10,* 1–17.

Wolf, L. F., & Smith, J. K. (1995). The consequence of consequence: Motivation, anxiety, and test performance. *Applied Measurement in Education, 8,* 227–242.

Wolf, L. F., Smith, J. K., & Birnbaum, M. E. (1995). Consequence of performance, test motivation, and mentally taxing items. *Applied Measurement in Education, 8,* 341–351.

Appendix

Rubric for Grading the Content of an Essay

Score	No. of theories	Criteria for evaluation
0	0	No content (word *motivation* does not count).
1	0	Several relevant terms, not explained or used inappropriately.
1.5	1	One or two theories mentioned appropriately, but the description is not full or confused.
2	1	One theory explained; other terms are used inappropriately or too lightly.
2.5	1	One theory well-explained; others are touched upon correctly (terms mentioned).
3	2	Two theories explained, but with some confused application, not enough detail and examples (some other theories may be touched on).
3.5	2	Two theories explained; description of one not full or confused (some other theories may be touched upon).
4	2	Two theories well explained, or terms from one or more theories mentioned.
4.5	2	Level 4 plus argument leading very well to conclusion.
5	3+	Three or more theories explained and properly applied, but with some confused terms and not enough detail for one of them.
5.5	3+	Three or more discussed theories, well explained and properly applied, with minor omissions.
6	3+	Three or more discussed theories, well explained, properly applied and substantiated by examples; other class readings are included.

Received November 17, 2008
Revision received September 9, 2009
Accepted September 21, 2009 ■