CHAPTER 12

FORMATIVE ASSESSMENT IN HIGHER EDUCATION

Frequency and Consequence

Jeffrey K. Smith and Anastasiya A. Lipnevich

INTRODUCTION

It is an unfortunate truism that how we teach at the university level is informed more by how *we* were taught at the university level than through any systematic examination of what is best for those whom we are teaching. We know full well the benefits of systematic feedback that can be delivered through formative assessment (Angelo & Cross, 1993; Nicol & Macfarlane-Dick, 2006), and yet we persist in lecture and assessment practices that differ little from what might have been seen a hundred years ago. The purpose of this chapter is to look at why this situation exists, to contrast assessment practices at the tertiary level to practices at the primary and secondary level, and to consider what might occur if we modified our practices to bring them more in line with what we understand to be good assessment practice.

To begin, there are certainly exceptions to this broad characterization of assessment practice at the tertiary level. Not only are there faculty who engage in exemplary instructional assessment practice, there are also fac-

Student Perspectives on Assessment: What Students Can Tell Us about Assessment for Learning,
pages 279–296

ulty who lean to a more time-honored and revered approach to teaching at the university level who are remarkably effective as instructors. There are many routes to effective university instruction. We all have our versions of "Mr. Chips," who brilliantly brought us through Econ 101, or Shakespeare, or even, perhaps in rare occasions, Organic Chemistry. The goal here is not to be the scold to faculty who take a traditional approach to university lecturing, but rather to explore the nature and structure of university life and instruction, and to see how that structure creates an environmental press toward assessment practices that are ill-aligned with what the scholarship in the field clearly recommends as best practice.

On one side of the equation, we have university faculty for whom teaching is a larger or smaller part of their professional identity and of their daily routine. Faculty range in attitude toward teaching from those who see it as the essence of their life's work to those who believe that the university would be a fine place if we could just rid ourselves of those pesky students. In addition to the personal dispositions of faculty toward instruction and assessment, there are structural constraints, particularly at research universities—as opposed to colleges and universities dedicated to teaching—that make moving on from a lecture/test mode of instruction difficult, if not impossible. The reward structure at research universities is based on research and scholarly accomplishments, and not on achievements in teaching. The person teaching Biology 101 in such institutions typically has a professional identity as a research biologist, not as a teacher of large groups of eighteen-year-olds. This may be beneficial to society in many ways, but exemplary teaching is not one of those benefits.

On the other side of the equation we have students. Students vary in their knowledge and skills coming into university, as well as in their self-regulatory abilities and orientation toward the academic work associated with higher education. The shift from the nature of instruction at the secondary school level to the university level can be dramatic and difficult to adjust to for many students. An anecdote might be appropriate here. A first year student we know at an ivy league school called his father with the following question: "Dad, I got one of those calendar things you told me to get, but I have a question on it. If I get an assignment, say, on a Thursday, and it's due say, a week from the following Tuesday, do I write that down on the Thursday or on the following Tuesday?" This was a National Merit Scholar, but he was accustomed to an instructional system from his high school where assignments were either made on a day-to-day basis, or, if longer term, had a teacher reminding him of due dates. Long-term planning and organization were not part of the curriculum, manifest or latent. Furthermore, as the same student put it upon graduating, "The thing that really surprised me about (this Ivy League University) was that I thought the instructors would be much better than my high school instructors, but in fact, they were nowhere near as good."

So we find a situation where on one hand, the university faculty may not have a strong incentive to focus on instructional issues, and on the other hand, students who are often not ready for a major shift in terms of their responsibilities as learners. Our goal here is to explore that disconnect, first by looking at the assessment practices of college instructors as revealed by an examination of their syllabi, and second, to explore what might happen if faculty gave students the benefit of detailed formative assessment feedback to guide their work. Both of these studies come from an inherently American perspective on tertiary education.

WHAT IS GOOD ASSESSMENT PRACTICE?

Roughly forty years ago, Michael Scriven (1967) developed the notion that evaluation could be formative or summative in nature; that is, evaluation could be focused on trying to improve instructional programs (formative), or trying to determine the overall quality of a program (summative). This distinction found its way fairly quickly to the student of assessment of student progress, perhaps most notably in Bloom's mastery learning instructional model (Bloom, 1968). As conceptualized by Bloom, formative assessment is essentially a form of feedback to students about their progress. This conceptualization is salutary in two fashions: first, it clarifies the essence of the distinction between formative and summative assessment, and second, it allows research on formative assessment to make use of the theoretical and research advances in the broader field of the effects of feedback. Black and Wiliam (1998) present a straightforward definition of feedback by arguing that the essence of formative assessment is the presentation of information regarding a learner's current state of knowledge and the desired state of knowledge.

There have been a number of reviews of literature in recent years that have looked at the impact of formative assessment and feedback on student growth. Crooks (1988) reviewed the field of evaluation of student progress broadly, drawing a number of conclusions about both the cognitive and affective impacts of good evaluation practice. Bangert-Drowns, Kulik, and Morgan (1991) conducted a meta analysis of the research on feedback and found that feedback, although generally quite effective, was not uniformly so. They concluded that in order to be effective, feedback had to promote the mindful engagement of students in the learning task. In a singularly impressive meta analysis of the effects of feedback, Hattie and Timperley (2007) came to similar conclusions as Bangert-Drowns, Kulik, and Morgan with regard to the need for feedback to engage the learner actively in trying to process the pertinent information to be learned.

If effective feedback causes students to engage in mindful and purposeful consideration of their strengths and weaknesses related to the learning task, what are the characteristics of formative assessment that enhance and promote such mindful engagement? The characteristics of high quality formative assessment are fairly well-understood (Stiggins, 2004, 2005):

- Accuracy: Feedback should be accurate in terms of it being a correct assessment of the learner's current state.
- Relevance: Feedback should be focused on the aspects of learning that currently need attention from the learner.
- Timeliness: Feedback should be received in a timely fashion in order for the learner to be able to properly relate it to his or her learning.
- Mediation: Frequently, learners need expert assistance in interpreting feedback and determining how to act upon it.
- Context: Formative assessment functions best in a setting that is supportive in nature and encourages learning from mistakes. Within school classrooms, this is often referred to as the "classroom assessment climate."

WHERE DO WE SEE GOOD ASSESSMENT? A "COOK'S TOUR" OF ASSESSMENT ENVIRONMENTS FROM LOWER TO HIGHER EDUCATION

If we know what good assessment practice looks like, where do we typically see it in educational settings, and why do we see it there? The literature on assessment in primary and secondary education might be characterized as split between research that focuses on technical aspects of measurement, and research that focuses on the processes of assessment that enhance instruction. There is far less assessment research that is focused at the tertiary level, and little of it looks at instructional processes (Orrell, 2006), which is too bad, because there is much that could be learned by those working at the tertiary level from looking at what is known from research at the primary and secondary levels. The literature that exists at the tertiary level focuses primarily on issues such as selection and placement into different levels of instruction, along with some work on competency issues with regard to various courses of study. There is very little empirical research on formative assessment in university teaching, and how such assessment might enhance the teaching/learning process. In part, this is because this is simply not an issue of critical importance at the tertiary level. But might it be? Might it be possible to see the characteristics of good assessment practice at the university level? What are the structural and environmental pressures that encourage or discourage good assessment practices?

Primary Instruction

Primary classrooms differ from university classrooms in many highly revealing ways. To begin, they house the same students for six or so hours a day for somewhere around 180 days in a year. Thus, the same group of students is with a teacher for somewhere around 1000 hours. The classroom is where they "live" for a great part of the year. It typically includes a space that students consider to be their own. They are taught by the same teacher for most of the 1000 hours, a teacher who is responsible for their academic as well as their overall physical and psychological well-being over the course of a school year. What are the assessment consequences of such a setting?

Such settings allow for a sense of trust to be built between student and teacher, for teachers to gain a broad and deep knowledge of student achievement and growth, and provide the time required for assessment information to be based on many sources of information and to be rich in its nature. Additionally, assessment feedback can presented in a timely fashion and can be tailored to the needs of individual students.

Secondary Instruction

Classrooms at the secondary level differ from instruction at both the primary and tertiary levels. Students typically will spend 40–50 minutes in a subject area every day for five days a week with the same teacher. Although this does not allow the same total exposure to the same teacher as in the primary classroom, there is still roughly 120–160 hours of instructional time spent with the same teacher, allowing for the development of some degree of personal relationship with students. But at the secondary level, there is the requirement in American schools for teachers to assign grades in courses, and these grades are important in highly competitive university admissions decisions. Thus, high school teachers play the role of the advocate of the student as well as the judge of the student, and these roles often come into conflict (Smith, Smith, & DeLisi, 2001). What are the assessment consequences in this general type of setting?

Secondary teachers can develop good assessment environments in their classes. They see students on a regular basis and can focus their efforts on a subject area in which they typically possess strong expertise. They do, however, work under several constraints. To begin, there is far less contact between teacher and student than in primary classrooms. A second constraint is that that the secondary teacher has many more students to work with—sometimes four to five times as many students as primary school teachers.

Tertiary Instruction

Instruction at the tertiary level varies widely, from small seminars to exceptionally large introductory lecture courses. A common element among these settings is the amount of contact hours between the instructor and the student, which is far less than is seen at the primary or secondary level. Furthermore, few institutions have a strong culture of instructional practice overall, much less assessment practice. Finally, as at the high school level, there is a strong emphasis on summative achievement, with the assignation of grades being the dominant assessment requirement. What are the assessment consequences of such a setting?

There is often little opportunity to provide formative assessment feedback at the tertiary level. There is too little time, too many students, and frequently too strong a demand to engage in summative assessment by giving grades to students. Furthermore, unlike many primary and secondary institutions, there is typically no notion of what standard assessment practice, or even instructional practice is at the tertiary level. There is great variation from department to department, and even from faculty member to faculty member within a department.

Structural Issues Constraining Good Assessment Practice

What can be seen in the description above is that there are many substantial constraints militating against good formative assessment practice at the tertiary level. In addition to the major constraints of time and the number of students, there is simply not a history of engaging in formative assessment at the university level. Most university faculty appear to use examinations and term papers to assess student performance. But what really is the case with regard to assessment practice at the tertiary level? Smith (2002) conducted a study of university course syllabi to look at teaching with technology, but the data set also can be used to reflect on the question of the degree to which university faculty engage in formative assessment practices, the first of the two fundamental questions under consideration in this chapter. The study and its results are summarized here.

STUDY 1: HOW COLLEGE FACULTY ASSESS STUDENTS

In an effort to look at issues involved in teaching using the Internet, this study examined a set of course syllabi from a large, eastern, state university. The primary issue was to see how much these courses relied on in-class tests, quizzes, or exams in the grading requirements of the course. The syllabi were examined for the percentage of the grading requirements that consisted of in-class testing. The data set was re-examined for purposes of this chapter to see how strongly the syllabi communicated expectations to students, and the degree to which there was indication of the use of a formative assessment perspective in the course.

The Sample of Syllabi

One hundred syllabi were selected from the publicly-available syllabi on the University's web-based course program. These are not distance-learning or electronically-delivered courses, they simply are courses that have a website for the course. There is something of a bias incurred here as more technologically-oriented faculty members are more likely to have a course website. The sample was stratified in two fashions, subject area and level. The four different subject areas chosen were:

1. Humanities
2. Social science
3. Science/math
4. Professional schools

The different levels of instruction were:

1. First and second year classes
2. Third and fourth year classes
3. Graduate classes

This stratification generated a three by four design. Although there were classes in each of the cells, the proportions were far from equal. In particular it was difficult to find classes at the graduate level outside of the professional schools, and, it was a little bit harder to find humanities courses than courses in other areas.

Measures

Each of the 100 syllabi was examined to determine what the grading system consisted of for the course. The first aspect of the syllabi that was apparent was the wide variety in the degree to which faculty choose to communicate to their students with regard to assessment. Some faculty made extensive comments about the goals or objectives of the course, and what would be expected of students. Other faculty provided no information other than what and when the exams or other assessments would be. Syllabi were rated on the following three-point scale:

3. Extensive communication about expectations;
2. Moderate level of communication about expectation; and
1. Little or no communication about expectations

Two researchers independently rated the syllabi, achieving an 88% initial agreement, and reaching consensus on the final ratings.

Next, a single number representing the percentage of the grade that was dependent upon in-class examinations was generated for each course in the sample. For about 10 of the 100 classes, it was necessary to estimate this percentage as the numbers weren't clear from the information in the syllabus (e.g., "….and you can raise your grade by as much as half a grade for truly exceptional classroom participation"). For 17 of the 100 courses, it was simply not possible to determine these percentages; in most cases, a grading system wasn't mentioned in the syllabus.

Finally, syllabi were rated on the degree to which they indicated that some level of formative assessment would be used in the course. This might take the form of a practice examination being made available, reviews for examinations being scheduled into the classes, opportunities to resubmit work after revision, or quizzes that would not be counted toward a final grade. Again, a three-point scale was used:

3. Extensive formative assessment;
2. Some clear indication of formative assessment; and
1 No indication of formative assessment (including 17 courses where no mention of assessment at all was made).

The same two researchers rated the syllabi for this variable independently, achieving a 79% initial agreement and reaching consensus on all final ratings.

Results

We begin by looking at the degree of communication to students. Overall, the ratings on communication were as follows:

- 22% Level 3, extensive communication
- 48% Level 2, moderate communication
- 30% Level 1, little or no communication

Results by level and by discipline were tested with a chi square procedure and found to be not statistically significant, but the individual differences were quite interesting. Some courses at the introductory (freshman or sophomore level) provided extensive information about what is expected of students and communicate well the nature of the course; others at the same level provide almost no information. The disparity between syllabi, when compared side-by-side, is striking. These syllabi represent courses often taken by the same students; some communicate clearly and with empathy toward students, others not at all.

The second analysis looked at the use of tests, exams, and quizzes in the course. Although this analysis does not speak directly to the question of the formative use of assessment, it is quite revealing concerning the mode of assessment that is typically employed. The raw distribution of the percentages of each course that was devoted to in-class quizzes, tests, or exams is presented in a stem and leaf diagram in Figure 1. What can be seen here is that the distribution is fundamentally tri-modal. Roughly one quarter of the courses rely exclusively, or nearly so, on tests, and roughly one quarter don't use them at all. The other half uses a combination of tests and other assessment tools. It should be noted that the courses using 90% or 95% testing almost all have the remainder of the grade points going to attendance, or class participation.

The mean usage of tests was 55%, with a median of 65%. Two analyses of variance were run to look at differences by subject matter area and by level of the course. It was originally planned to run this as one ANOVA, but at the graduate level, the preponderance of syllabi available were in professional schools, so one ANOVA was run with all 83 syllabi to look at subject area effects, and a second was run with the graduate level excluded to look

9	0005
8	0005
7	000000555555
6	000000557
5	000000
4	0
3	05
2	000
1	0055
0	00000000000000000000

FIGURE 1. Stem and Leaf Diagram of the Percentage of the Course Grade that Is a Quiz, Test, or Exam

TABLE 1. Means and Standard Deviations of Percentage of Tests Used in Grades

	Subject Area							
	Humanities		Science/Math		Social Science		Professional	
Level	M	SD	M	SD	M	SD	M	SD
Frosh/ Sophomore	29.0	36.0	74.3	36.1	78.6	18.0	69.0	41.0
Junior/ Senior	53.0	30.3	85.4	25.1	53.3	50.3	51.9	36.8
Graduate*	0.0	0.0	40.0	0.0	0.0	0.0	32.5	34.8

Note: *At the graduate level, there were only 2 syllabi for Humanities, and 1 each for Science/Math and Social Science.

at level effects. Significant differences were found for both subject area and level ($p < .05$), but no interactions were found. Means and standard deviations are presented in Table 1. Using a Scheffé *post hoc* analysis, we found that higher levels (junior/senior) use testing more than lower levels (freshman/sophomore). The humanities test less than any other group, and the sciences test more.

The final analysis looked at the degree to which the use of formative assessment was indicated from an analysis of the syllabi. The results here involve seem to support the contention that formative assessment simply is not a large component of tertiary level instruction. The results by level were:

- 6% Level 3, extensive formative assessment;
- 12% Level 2, some clear indication of formative assessment; and
- 82% Level 3, no indication of formative assessment (including 17 courses where no mention of assessment at all was made).

At first glance, it seems like there is very little concern for formative assessment and feedback at the tertiary level (in this institution). However, it needs to be pointed out that this analysis rests solely on what is presented in the course syllabus. It may well be the case that some courses did engage in formative assessment, but did not mention it in the course syllabus.

Discussion of Study 1

The analyses of the syllabi suggest a number of interesting patterns in university level instruction. It should be pointed out at the beginning that what is listed in a syllabus and what occurs in a course may not be the same thing. However, it has been our experience that there is a reasonably good correspondence between the two as the course syllabus is often viewed as a kind of agreement between the learner and the instructor. So we draw conclusions here with a level of tentativeness, but with some expectation that we are not too far off base.

First, we see that the level of communication about expectations for students varies substantially, with most courses providing at least a moderate level of expectation. However, 30% provide little or no advice to students, many of whom are first year students making a shift from a setting where communication is typically quite strong. This change may be quite difficult for some first year students to make, and may contribute to the severe problems that many first year students encounter when entering university study. Next, we see that there is a very wide variety of approaches to course assessment in general. Some courses rely exclusively on examinations, tests, and quizzes for marking, and others do not use them at all. As might be expected, usage varies by discipline and level, but not always in anticipated ways. It was interesting to see more testing going on at the higher level than the lower level. This appears to be primarily due to freshman/sophomore English courses having a large writing component, whereas the upper level courses are more literature-based and assessed by tests instead of written assignments. Again, as expected, there is less testing in humanities courses, more testing in science and math courses, with the social sciences and professional schools somewhere in the middle. Finally, we see very little evidence of assessment for formative purposes being used in any of the courses (based on the information in the syllabi).

Given the constraints of time, frequently very large class sizes, and, at research institutions, a stronger press for publication than high quality instruction, it is not surprising that evidence of good formative assessment evidence is scant based on an examination of course syllabi. This brings us to the second question pursued in this chapter: What kind of effectiveness might a formative assessment approach have if it were undertaken in a college course? To look at this question, we turn to a study conducted to look at aspects of formative assessment in a large, first-year undergraduate psychology course.

STUDY 2: THE EFFECTS OF FORMATIVE ASSESSMENT IN A LARGE UNDERGRADUATE COURSE

This second study was a large, *in situ* investigation into the effects of different aspects of assessment feedback. It was conducted at two eastern, state universities in introductory psychology classes taught by the same professor. The results for the two classes were quite similar and thus were combined into one data set, which is reported on here. The study was designed to look at the effects of detailed feedback on an essay examination that allowed for revision of that essay based on the feedback before a final grade was assigned. An extensive report of the findings can be found in Lipnevich and Smith (in press a) and Lipnevich and Smith (in press b).

Sample

The sample for the study was taken from two universities in the eastern United States where an introductory psychology course was being taught at both institutions by the same faculty member. There were 409 students in one university, and 55 in the other. The characteristics of the students at the two universities were quite similar, as were the final results from both groups, so they were combined into a single sample for purposes of reporting. The classes were predominantly made up of first and second year students, with a mean age of 18.9 ($SD = 2.5$); 52% of the participants were women. Roughly half of the participants self-identified as Caucasian, with roughly one quarter Asian, and the rest Hispanic, African American and other groups. Roughly 80% of the students in the sample were native English speakers.

Measures

A variety of measures were used in the study. The primary ones reported on here are the numerical score students received on their initial draft of their essay and their final score after revision on the essay. Prior to writing their initial essay, participants were given a measure of performance versus mastery goal orientation, The Learning and Performance Goal Orientation Scale, which was adapted for this research from (Button, Mathieu, & Zajac, 1996). The draft scores were a combination of scores from an electronic scoring program, E-Rater, developed by Educational Testing Service (Attali & Burstein, 2006; Burstein, 2003), and by the researcher in the study and the course professor. E-Rater provided scores and feedback on the more technical aspects of the writing and the researcher and professor provided scores and feedback based more on content. Scores were determined by rating each essay on a 0–6 scale for mechanics (by E-Rater) and content (by the professor and the researcher). A 30% weighting was given to the mechanics, and a 70% weighting to the content. Scores were then transformed to a 100-point scale. On a subsample of the papers, a 95% agreement was found between the two human raters.

Procedure

The study began with an essay test based on the first section of the course (motivation theory). Students wrote the essay on computer, and if they chose to participate in the experiment, completed a questionnaire, and were told to come back the following week to complete the experiment. In the second session, they were randomly assigned to one of several feedback conditions, and allowed to revise their essay. They then completed several more questionnaires concerning their reactions to the essay and the feedback they received. Their final grade for the exam was based on their second essay. (Students were allowed to decline participation and still receive feedback and be allowed to revise their essays, and final scores for all students were adjusted according to mean differences in groups so that all students would benefit as much as the students in what turned out to be the most advantageous group).

Students were assigned to three different feedback conditions: no detailed feedback, detailed feedback that was said to come from the instructor, and detailed feedback that was said to have been computer-generated. (In the rest of the discussion of this study, these will be referred to as "no feedback," "instructor feedback," and "computer feedback"). All feedback in fact was generated in the same fashion, as described below. This factor was crossed with two other factors. The first was whether students received a tentative initial numerical score (on a scale of 0–100), and whether a student received a statement of praise and encouragement that was tailored to their initial level of performance. This resulted in a 3 × 2 × 2 design. The score on the initial performance was used to block students into groups and as a covariate in the analysis. Using the blocking variable, students were randomly assigned to conditions. The final score on the essay counted as part of their grade in the course, so the exam was consequential for all students, an important factor on a task that required engagement and effort (Wolf, Smith, & Birnbaum, 1995).

Students in both feedback conditions received information concerning their performance both at the level of the grammar and writing style, and at the level of conceptual issues reflecting the content of their essay. Students receiving detailed feedback were informed on computer that their feedback was either generated by a computer-scoring program (computer feedback) or came from their instructor (instructor feedback). If they were in the grade condition, they also received a tentative initial grade. If they were in the praise condition, they received one of three praise statements, depending on how well they had done on their initial draft (e.g., "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.").

Results

The results of the study strongly support the notion that constructive, detailed feedback can enhance student performance. A 3 × 2 × 2 analysis of covariance (with a Bonferroni adjustment made for multiple comparisons) yielded a strong main effect for the feedback factor (computer feedback, instructor feedback, no feedback) F (2, 450) = 69.23, $p < .001$, $\eta^2 = .24$, a small but significant main effect for grade F (1, 450) = 4.07, $p < .05$, $\eta^2 = .04$, and a small but significant interaction between grade and praise F (1, 450) = 6.00, $p < .05$, $\eta^2 = .04$. Although interaction terms are often discussed first in presenting results, it is simpler in this case to begin with the strong main effect for feedback. In the no feedback condition, the mean final draft score was 74.85 (SD = 8.49); in the instructor feedback condition, the mean was 81.82 (SD = 7.94), and in the computer feedback condition was 80.24 (SD = 8.18). We can see here that the effect size was quite large ($\eta^2 = .24$, with an effect size difference between the no feedback and instructor feedback conditions of .85). *Post hoc* analyses showed that the no feedback condition was significantly different from both the computer and instructor feedback conditions, and that these two feedback conditions did not significantly differ from one another.

The main effect for grade needs to be considered jointly with the significant grade by praise interaction. Students who received a tentative grade on their initial draft performed *less* well on the final draft (M = 78.63, SD = 8.15) than those who did not receive a tentative initial grade (M = 79.25, SD = 9.24). The effect size here is small (.07), but the direction is particularly interesting. An initial grade seems to have hindered performance, not helped it. The interaction effect shows that a statement of praise helped to ameliorate the effects of receiving an initial grade. Students who received no grade and no praise got the highest scores (M = 79.55, SD = 9.20). Students who received a grade, but no praise got the lowest scores (M = 76.80, SD = 8.02). Students who received a grade, but also got a statement of praise performed almost as well as the students who got no grade and no praise (M = 79.16, SD = 8.24). Students who received no grade, but did receive praise scored similarly to the grade/praise group (M = 78.95, SD = 9.32). Thus, what we see with regard to grades and praise is that grade seems to depress performance slightly, but a statement of praise in addition to a grade lessens the negative effect of the grade.

We broke the sample down into three subsamples based on initial draft score (low, medium, high) to see if results were substantially different according to level of initial performance. There were some slight differences, but none that were strongly different from the overall results. We also broke the results down according to whether students reported a mastery goal orientation toward learning in this course versus a performance goal orientation. Although it wasn't part of the purpose for collecting goal orientation, we found that many students had both orientations, while others seemed to have neither. Furthermore, the relationship between goal orientation and performance on the initial or final essays were non-significant.

At the end of the experiment, a number of focus groups were held to ensure that students understood the experimental conditions (they did) and to get more detail on how they responded to the conditions they were in. We learned that students were slightly less confident about the accuracy and usefulness of the computer-based feedback. We also learned, somewhat to our surprise, that students were overwhelmingly positive about the experience as a whole. They thought that the idea of trying a first draft, getting feedback, and then getting to revise the draft before a final grade was assigned was a terrific idea—one that they had not yet experienced in any other university class. Students who received a low initial grade reported being somewhat depressed and deflated by that grade, while students receiving a high initial grade sometimes reported that they either didn't want to risk too much in a revision, or didn't feel that they had to.

Discussion of Study 2

The basic finding here is simple: formative assessment works. It is difficult to draw any other conclusion from this study. To be sure, we see some other interesting things going on in the study, but the efficacy of feedback in helping students to improve their performance is undeniable. We are intrigued by several ancillary aspects of the study. Feedback thought to be from an instructor or from a computer had only marginally differential effect if any and was not statistically significant. Receiving an initial grade on an exam seemed to have a slightly negative effect on subsequent performance, either by discouraging students, or perhaps by making them feel like the return on extra effort would not be worth it.

Having stated rather strongly that formative assessment works, we must, in fairness, also report that providing the feedback to students on this essay examination took an incredible amount of work that had to be executed in a relatively short time span, even with the help of a computer essay scoring program. The constraints on providing good formative assessment information mentioned above—time, effort, number of students—were certainly realized here.

CONCLUSIONS

In higher education, we find ourselves in the interesting situation of not engaging in instructional practices that seem to be clearly in the best interests of our students' learning. The evidence presented here strongly suggests: 1) that this is clearly the case, and that 2) that students would learn more if we provided them with formative assessment feedback. There are many reasons that can be given to explain this situation ("....only so many hours in the day....," I have to do my research as well....," "When does it become the students' responsibility to take control of their own learning?", "Give me another two teaching assistants, and I'll get this done," etc.). We have noted that there are a number of structural constraints on faculty that militate against spending more time on their classes in general, much less something as time consuming as formative assessment activities.

What to do? Simply saying that formative assessment would be beneficial to students does not mean that it will or can be done. Have all we accomplished here is an academic version of "belling the cat"? Perhaps not. Perhaps there are recommendations that can be made based on the two studies discussed in this chapter. The first recommendation would have to do with commitment. Without a belief that something should be done and can be done, we will remain in the same situation we are in today. But there are steps that can be taken, and some of them are fairly small ones. We would recommend beginning by looking at what already exists within the higher education community, perhaps by considering a compendium of teaching tips for the tertiary level by Mckeachie and Svincki (2006), or by looking more specifically at assessment techniques (Angelo & Cross, 1993).

Technology also has strong potential for playing a role here, whether it is from a computer-based essay scoring program such as E-Rater, or from computerized "clicker" systems. These clicker systems allow a faculty member to take a reading of where a class is with regard to an issue mid-lecture by having them respond to a multiple choice question, or register their perceived level of understanding electronically. The response from the class can be displayed instantaneously via computer projector. Another alternative is peer assessment of work, which can be helpful to both the assessor and the assessed. One might also consider the provision of exemplars of strong and weak work with regard to an assignment that lets students contrast their efforts to standards to see how they are doing. Finally, one might simply consider the benefits of clearly stating to students what is to be expected of them. Surely in most courses, one of the course objectives is *not* that the students can divine the instructional intentions of the professor.

In the end, the question with regard to formative assessment is not so much: Will it work? as it is: Do we have the commitment to use something that we know does work? This commitment can be, and is, undertaken by individual faculty who believe strongly in their responsibilities as teachers. But substantial constraints exist on faculty members, constraints that can be alleviated by institutions committed to quality instruction. We can work individually for the growth of the students that we encounter as faculty, and

we can work collectively for the betterment of our universities as institutions of higher learning.

REFERENCES

Angelo, T. A. & Cross, K. P. (1993). *Classroom Assessment Techniques: A handbook for faculty* (2nd ed.). San Francisco: Jossey-Bass.

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater v.2. *Journal of Technology, Learning, and Assessment, 4*(3), 123–212.

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.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7–68.

Bloom, B. S. (1968). Learning for mastery. *Evaluation Comment, 1*(2), 1–12.

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.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Button, S. B., Mathieu, J. E., & Zajac, D. M. (1996). Goal orientation in organizational research: A conceptual and empirical foundation. *Organizational Behavior and Human Decision Processes, 67*(1), 26–48.

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

Hattie, J. & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112.

Lipnevich, A. A., & Smith, J. K. (in press a). The effects of differential feedback on students' examination performance. *Journal of Experimental Psychology: Applied*

Lipnevich, A. A., & Smith, J. K. (in press b). "I really need feedback to learn": Students' perspectives on the effectiveness of the differential feedback messages. *Educational Assessment, Evaluation and Accountability.*

McKeachie, W. J., & Svinicki, M. (2006). *McKeachie's teaching tips: Strategies, research, and theory for college and university teachers (12th ed.).* Florence, KY: Cengage Learning.

Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education, 31*(2), 199–218.

Orrell, J. (2006). Feedback on learning achievement: rhetoric and reality. *Teaching in Higher Education, 11*(4), 441–456.

Scriven, M. (1967) The methodology of evaluation. In R. E. Stake (Ed.), *AERA monograph series on curriculum evaluation* (No. 1). Chicago: Rand McNally.

Smith, J. K. (2002). "Challenges for assessment in technology-based instruction in higher education." Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Smith, J. K., Smith, L.F., & De Lisi, R. (2001). *Natural classroom assessment.* Thousand Oaks, CA: Corwin Press.

Stiggins, R. J. (2004). New assessment beliefs for a new school mission. *Phi Delta Kappan, 86,* 22–27.

Stiggins, R. J. (2005). *Student-involved assessment for learning (4 Ed.).* Upper Saddle River, NJ: Pearson Prentice Hall.

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.