Implications of Fail-forward in an Online Environment under Alternative Grading Schemes

Hilde Patron, University of West Georgia in Carrollton

William J. Smith, University of West Georgia in Carrollton

Abstract

The concept of fail-forward can be used as a teaching technique to motivate students to learn from their mistakes. For example, when students are allowed to re-work incorrect responses on a test for a partial grade they are failing-forward. In this paper we look at the effects of failing-forward on student effort in online learning environments. We consider two alternative grading schemes with different levels of penalties for failure. Under the first grading scheme students are allowed to submit their work up to three times without being penalized. Under the alternative grading scheme students are penalized every time that they answer questions incorrectly. We find that instructors may be able to use the “average of all scores” grading scheme to increase the level of preparation of students even with differences in students’ innate ability. However, the benefits are less pronounced in fully online classes where there is no face-to-face instruction time.

Keywords: fail-forward, student performance, online learning, business statistics

Introduction

In the classroom environment, it is common for instructors to posit a question with the primary intention of initially eliciting incorrect responses from students. The process of first unearthing student's deeply rooted preconceived notions, and then allowing the student to discover the better answer is a well-established pedagogical tool, which has become known as *failing-forward*. However, questions remain about how the assignment of credit during such an exercise may influence a student's incentives to study before the exercise. If a student is allowed unlimited attempts at a problem and receives full credit once the correct answer is provided, there would be little incentive for the student to do much more than guess repeatedly. To varying degrees, fail-forward is used as a teaching/learning technique from the time each of us is able to explore the world around us. The disappointment and sometimes physical pain involved in making mistakes is often used by parents and teachers to crystallize important concepts in the learning mind. We focus on a specific incarnation of fail-forward in business statistics taught either fully or partially online to examine the way students respond to different levels of credit afforded to them on second (and even third) chances. Since fail-forward can be applied in a myriad of different ways, to get the most out of fail-forward as a pedagogical tool, it is important that we clearly understand how adjusting the credit given affects students' incentives to actually learn from their mistakes rather than just becoming efficient guessers. It is also important to understand how the benefits from fail-forward techniques differ between fully online courses and courses in which instructors can quickly correct students' misconceptions face-to-face.

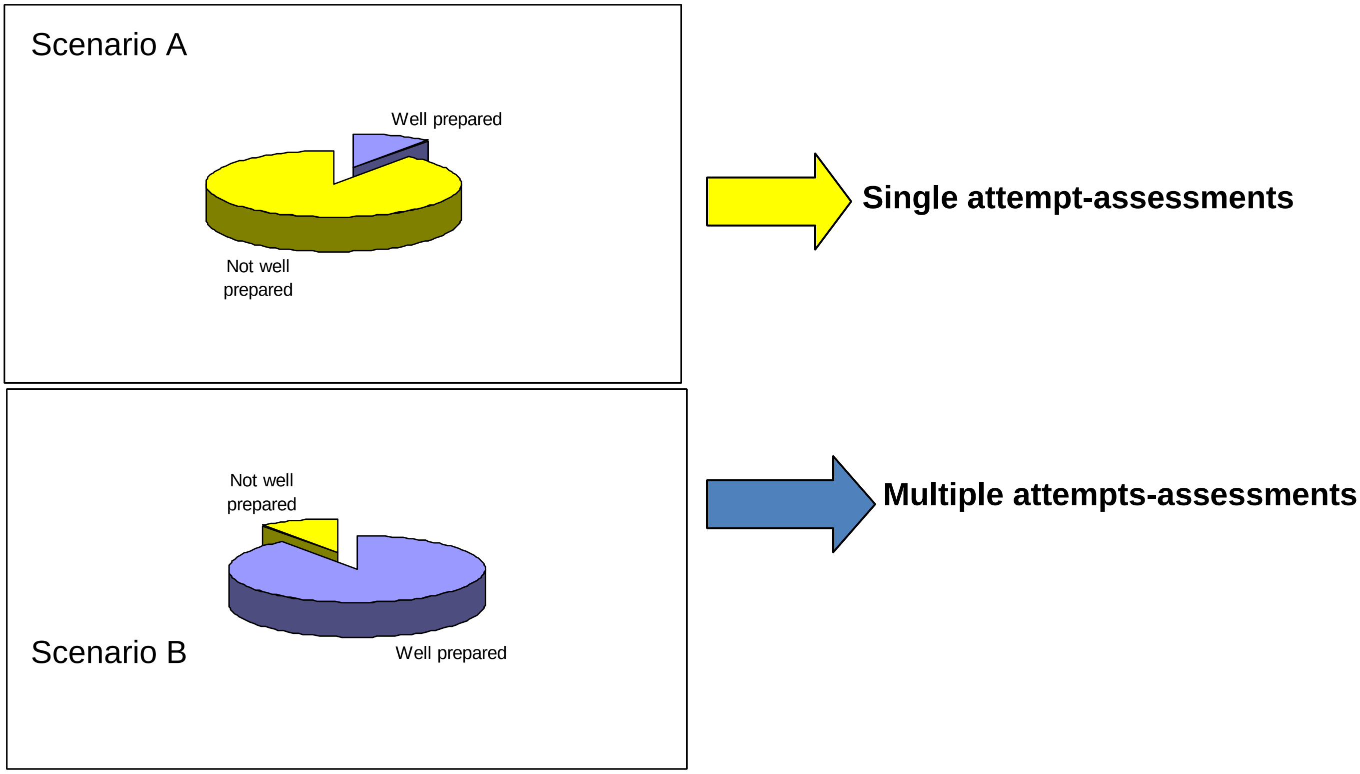
The benefits of fail-forward techniques can be meaningful; students can improve their grades and can learn what they failed to learn in the first place. Instructors have to be careful, however, in the way they respond to the students who make mistakes. They have to make sure to give students useful comments and not feedback that discourages thinking.[1] Instructors also have to be wary of the way in which they administer fail-forward techniques. If students are able to search for responses, guess at their answers, or otherwise "game the system" their learning can be impaired by these methods.

[1] See, for example, Salomon & Globerson (1987) and Bangert-Drowns, Kulik, Kulik, & Morgan (1991).

The empirical evidence regarding the benefits of fail-forward techniques is inconclusive. For instance, some authors find evidence in support of the hypothesis that allowing students to redo their work on multiple occasions increases student learning or performance, including Bangert-Drowns, Kulik, Kulik & Morgan (1991), Coates & Humphreys (2001), and Patron & Smith (2009). However, in a Meta Analysis study Clariana (1993) finds that while “repeat until correct” or “multiple attempt assignments” are superior to no-feedback assignments, in some cases “repeat until correct” feedback is inferior to “single response feedback”. Kulhavy (1977), Clariana & Smith (1989), and Clariana (1990) further find that the optimal type of feedback depends on the students’ prior knowledge and beliefs. For example, feedback that corrects misconceptions is more valuable than feedback that reinforces accuracy. In addition, multiple-attempt assessments with feedback are better for students with “high prior knowledge,” e.g., students who have read and studied the material prior to the exercise; however, single attempts with correct response feedback are better for students with “low prior knowledge,” possibly because it increases the incentive for students to become more familiar with the material prior to the exercise.

The extant literature suggests that students can benefit from fail-forward techniques when they approach their assessments well prepared (or with “high knowledge”); students who are not well prepared are better off (learning wise) with single attempt assessments (see, e.g., Kulhavy 1977, Clariana & Smith 1989, and Clariana 1990). Figure 1 summarizes this idea: an instructor faced with students who are mostly unprepared should not use fail-forward techniques, whereas an instructor with a class made mostly of well-prepared students should use fail-forward techniques.

FIGURE 1: Class Distribution and Optimal Type of Assessments

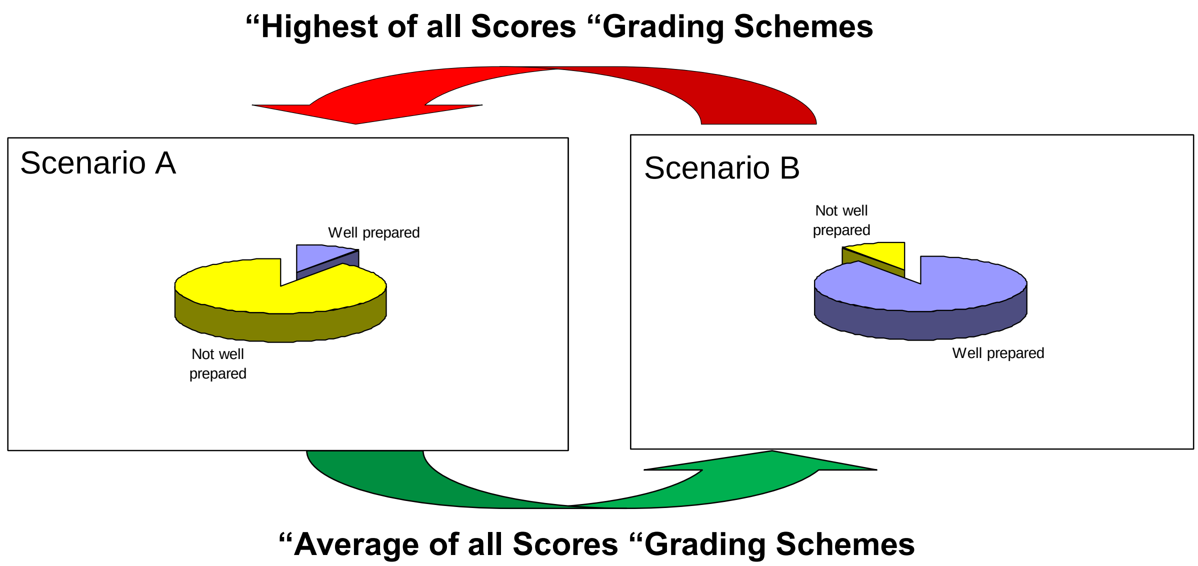


In this paper we look into some options that instructors have to manipulate the distribution of well-prepared versus not-well-prepared students in their courses. More specifically, we study whether or not instructors can manipulate students' incentives within the fail-forward framework by using alternative grading schemes in on-line assessments. Furthermore, we study how effective instructors of online courses are at manipulating incentives relative to instructors of hybrid courses. Our results suggest that by averaging each of the quiz attempts for the quiz grade, students likely respond by increasing their preparation prior to the first attempt. However, results are most pronounced in courses with some degree of face-to-face instruction.

Our data are obtained from quizzes from students registered in introduction to business statistics courses at a mid-size public university. Students in multiple sections of the course were give ten quizzes during the semester and were allowed to take each quiz up to three times. Students were divided into two groups. One group was allowed to complete quizzes without a penalty; the highest score was assigned as the quiz score. The other group faced a cost when answering questions incorrectly; the average score of all attempts was assigned as the quiz score. The aim

of this study is to determine whether or not the incentives created by the two grading schemes have an impact on student effort or preparation for quizzes. More specifically, we suspect that by assigning grades using the costly "average-of-all-scores" grading scheme instructors can shift the distribution where a larger share of students engage in a "high level of preparation" as opposed to "low level of preparation" shown in Figure 2.

FIGURE 2: Manipulation of Student Incentives (via costs)



The rest of this paper is organized as follows. The next section describes the data and method of the study, and the following section presents results. The last section concludes and discusses avenues for further research.

Description of the Study

Students registered in two fully online sections and two hybrid (partially online) sections of introductory business statistics during the fall and spring semesters of 2009 were given ten online multiple choice quizzes during the semester. There were a total of 135 students registered in the four sections of the course. Sixty two were registered in fully online sections and 73 in hybrid courses. Quizzes were delivered online using the website *Statsportal*. Each student was allowed to complete the quiz up to three times. After each submission the student learned his or her total

score and which questions were answered incorrectly. Questions remained the same between attempts. Roughly half of the students (those taking the course in the spring semester) were assigned the highest score of the three attempts as the quiz score. Under this approach students were not penalized for answering questions incorrectly in the first two attempts. The rest of the students (those registered in the fall) were assigned the average of all scores as their quiz score.[2] This harsher grading method penalized students for answering questions incorrectly in the first two attempts.

We use quiz scores to analyze the effect of the two grading schemes on student effort or preparation. We approximate effort with the score on the first attempt at each quiz. We calculate t-tests of differences of means and we run regression models to estimate the determinants of the "first attempt score" at the quiz. The variables that we include in the analysis and summary statistics are presented in Table 1. The grades in the quizzes were obtained from instructors' records. All other information was obtained from the university's office of the registrar.[3]

In the regression models the unit of study is quizzes. More specifically, the dependent variable is the percent of correct responses each student earned in each of the ten quizzes. Since there were 135 students registered at the beginning of the semester and each student was asked to complete ten quizzes we have potentially 1,350 observations in the regressions, 620 in online course regressions and 730 in hybrid course regressions. However, since not all students completed all quizzes the number of observations in our estimations varies from these totals. We do our estimations following a two step process to allow for sample selection correction. Following Heckman (1979), we first estimate the individual probability that a student registers for the online class using a probit model. In this first step the dependent variable is a categorical variable that equals 1 if the student registers for a fully online class and equals 0 otherwise; the independent variables are the student's age, gender, cumulative GPA, and number of hours he/she is enrolled in. In a second stage we use a transformation of these predicted probabilities to estimate a linear model (using least squares) to estimate the first-attempt-quiz-scores. More

[2] More specifically, 67 students had their quizzes graded under the "average of all scores" grading scheme (29 in the online sections and 38 in the hybrid sections) and 68 (33 in online courses and 35 in hybrid courses) had their quizzes graded under the "highest of all scores" grading scheme.

[3] The data used in this study is part of data regularly corrected in the College of Business for evaluation and assessment of core courses.

specifically, the dependent variable is the quiz score and the regressors include the student GPA, the grading scheme dummy, dummy variables for each of the ten quizzes, and the transformation of the predicted probabilities obtained in the first estimation (also known as the Inverse Mills Ratio).[4] Variables used in the analysis and descriptive statistics are presented in Table 1.

TABLE 1: Description of Variables*

Variable	Description	Mean	Std.Dev.	Min.	Max.	Obs.
AGE	Student age in years.	23.76	6.43	19.00	53.00	1350
ENROLLED_HOURS	Number of hours the student is enrolled in.	13.53	2.85	3.00	20.00	1350
MALE	Dummy = 1 if gender = male.	0.47	0.50	0.00	1.00	1350
GPA	Student grade point average.	2.87	0.49	1.75	4.00	1340
QUIZ_FIRST_ATTEMPT	Score on the first attempt at the quiz.	59.03	24.97	0.00	100.00	1260
AVERAGE_DUMMY	Dummy variable that equals 1 if grading scheme is "average-of-all-scores".	0.50	0.50	0.00	1.00	1350
ONLINE_DUMMY	Dummy variable that equals 1 if student is taking class online.	0.46	0.50	0.00	1.00	1350
HYBRID_DUMMY	Dummy variable that equals 1 if student is registered in hybrid section.	0.54	0.50	0.00	1.00	1350

*We also use ten dummy variables, one for each quiz. For example, Quiz_1_DUMMY equals 1 for quiz # 1 and equals 0 otherwise.

Results

Table 2 summarizes the average scores on the first attempt at the quizzes. The first panel of Table 2 includes all students, the second panel only includes students subject to the "average of all scores" grading scheme, and the third panel includes data for students subject to the "highest

[4] For a review of sample selection models and two stage estimations see Winship & Mare (1992).

of all scores" grading scheme. Finally, the fourth panel presents the results of t-tests for mean differences. The tests are calculated according to the formula $t = \frac{\bar{x}_{Highest} - \bar{x}_{Average}}{\sqrt{\frac{s^2_{Highest}}{n_{Highest}} + \frac{s^2_{Average}}{n_{Average}}}}$, where $\bar{x}$ denotes the mean quiz score, s^2 denotes the variance of the quiz score, and n denotes the number of observations. The subscript "Highest" refers to the "highest of all scores" grading scheme and the subscript "Average" refers to the "average of all scores" grading scheme.

TABLE 2: Average Score on First Quiz Attempt

	All students			"Average of all scores" scheme			"Highest of all scores" scheme			t-tests		
	All	**Online**	**Hybrid**	**All**	**Online**	**Hybrid**	**All**	**Online**	**Hybrid**	**All**	**Online**	**Hybrid**
Quiz 1	75.71	77.20	74.41	77.55	80.00	75.70	73.90	74.83	72.96	-1.06	-1.62	-0.49
Quiz 2	64.41	62.25	66.17	65.12	63.26	66.45	63.66	61.33	65.83	-0.48	-0.42	-0.15
Quiz 3	56.54	56.16	56.85	59.58	59.48	59.66	53.49	53.36	53.61	-1.60	-0.98	-1.26
Quiz 4	66.14	63.77	68.23	69.51	63.93	73.69	62.87	63.64	62.08	-1.85*	-0.05	-2.42**
Quiz 5	62.85	62.86	62.84	68.40	68.37	68.42	57.31	58.22	56.37	-3.49***	-2.28**	-2.64***
Quiz 6	55.79	51.53	59.42	59.17	52.15	64.30	52.30	50.99	53.60	-1.98**	-0.22	-2.39**
Quiz 7	47.93	45.34	49.99	54.00	48.26	57.57	42.06	43.18	40.94	-2.76***	-0.75	-2.90***
Quiz 8	46.09	39.68	51.37	53.74	45.46	59.11	38.69	35.35	42.14	-3.15***	-1.42	-2.67***
Quiz 9	53.06	43.89	60.45	58.97	47.27	66.11	47.62	41.56	53.87	-2.25**	-0.70	-2.02**
Quiz 10	60.28	53.89	65.47	71.81	66.04	75.70	48.95	44.09	53.65	-4.03***	-2.62**	-2.88***

*** Significant at the 1% level, ** Significant at the 5% level, * Significant at the 10% level

The test statistics are all negative and many of them are statistically significant at traditional levels. This suggests that students who were awarded points based on the "average of all scores" grading scheme tended to prepare more intensively or expended more effort in answering questions correctly on their first attempts. This result is especially noteworthy in hybrid courses, where the mean differences are significant for 7 of the 10 quizzes. Interestingly, earlier quizzes are insignificant. This may be due to students becoming more familiar with the grading scheme and thus developing more effective quiz-taking behaviors as the course progresses. For the all-online courses, the students do not appear to adapt as quickly or as effectively, possibly due to the lower level in face-to-face interaction with other students in the class.

To properly study the impact of grading schemes on student effort however we have to control for student innate ability and other factors. First of all, because we want to allow for an individual's ability to self-select into the two different types of courses (all-online and hybrid), and allow for the course type to affect quiz outcomes, we estimate a two-stage model with sample selection based on Heckman's approach (1979). As mentioned above, we estimate the individual probability that a student registers for the online class as a function of the student's age, sex, prior cumulative grade point average, and hours enrolled in using a probit model. We then use a transformation of these estimated probabilities to estimate the first-attempt-quiz-scores using ordinary least squares. Although the non-linear form of the first stage model provides identification, we include further exclusion restrictions in the second stage as well. We exclude the age, sex and enrollment hours from the second stage because these variables were found to be correlated with the likelihood of enrolling in online classes, but not correlated with the scores on the first attempt for each quiz.

Table 3 summarizes the first step of the sample-selection estimation. The model suggests that online students are more likely to be older, female, have a lower GPA and enroll in fewer course-hours than students who take the course on campus. These findings are consistent with what has become the target market for online classes in universities, working-aged adults who are seeking part-time educational opportunities. Furthermore, each of the explanatory variables is significant at the 1 percent level.

TABLE 3: Step 1 of Sample Selection Estimation.

Dependent Variable: ONLINE_DUMMY				
Variable		**Coefficient**	**P-Value**	
AGE		0.07	0.00	
MALE		-0.36	0.00	
GPA		-0.27	0.00	
ENROLLED_HOURS		-0.06	0.00	
			Predictions	
		0	1	Total
	0	660	70	730
Actual	1	330	280	610
	Total	990	350	1340

The results from the second stage (least squares regression) are summarized in Table 4. We estimate two different models, one for students registered in online classes and one for students registered in hybrid classes. The table shows that after controlling for GPA, the "average of all scores" grading scheme leads to higher first attempt quiz scores. More specifically, first attempt scores are on average 7.31 points higher among students registered in online courses, and 11.46 points higher when only students in hybrid courses are considered, holding initial GPA constant. Both fully online and hybrid class students perform significantly and substantially better on the first attempt when every attempt counts toward the grade. This may be an indication of better preparation on the part of the student for the quiz, a more serious and focused first attempt, or both.

TABLE 4: Step 2 of Sample Selection Estimation

Dependent Variable: QUIZ_FIRST_ATTEMPT				
	Online courses		**Hybrid courses**	
Variable	**Coefficient**	**P-Value**	**Coefficient**	**P-Value**
Quiz_1_DUMMY	20.69***	0.00	36.30***	0.00
Quiz_2_DUMMY	5.66	0.37	27.87***	0.00
Quiz_3_DUMMY	0.46	0.94	18.73***	0.00
Quiz_4_DUMMY	7.58	0.22	30.21***	0.00
Quiz_5_DUMMY	6.93	0.26	24.65***	0.00
Quiz_6_DUMMY	-4.29	0.49	21.04***	0.00
Quiz_7_DUMMY	-11.36*	0.07	11.77*	0.07
Quiz_8_DUMMY	-16.83***	0.01	13.08**	0.04
Quiz_9_DUMMY	-11.60*	0.07	21.98***	0.00
Quiz_10_DUMMY	-2.77*	0.10	27.23***	0.00
GPA	17.65***	0.00	9.97***	0.00
AVERAGE_DUMMY	7.31***	0.00	11.46***	0.00
LAMBDA	3.47	0.18	4.96	0.17
Observations		564		686
Parameters		13		13
Sum of squares		240261.3		333774.3
R-squared=		0.32		0.18
Adjusted R-squared =		0.30		0.16
F[k-1,n-k] =		20.98		11.95
Log-L =		-2507.63		-3095.65
Restricted Log-L =		-2620.34		-3168.46

Students in hybrid classes perform substantially and statistically significantly better than their counterparts from fully online classes. This difference may be due to the benefits derived from being in a classroom environment with an instructor. The two coefficients may provide, in small part, a measure of the value added by the instructor in the classroom setting as opposed to the online setting. Stated another way, the limited in-class instruction and interaction in the hybrid course may account for an additional 4.15 points on average for the first attempt on quizzes, or almost a half letter grade. Furthermore, GPA is significant for both online and hybrid courses; however, GPA has a substantially larger effect for the fully online students. This is an indication that students with higher academic abilities in either class setting perform better on their first attempts; however, comparing the coefficient from both models indicates innate ability is more important for first quiz attempts when in-class, face-to-face instruction is more limited, as it is with the fully online classes.

Taken as a whole, our results suggest that instructors may be able to use the "average of all scores" grading scheme to increase the level of preparation of students even with differences in students' innate ability. However, the benefits are more pronounced for classes where there is at least some face-to-face instruction time. Furthermore, students with higher innate academic ability appear to be the students who recognize and respond most effectively to the more punitive grading scheme. These students increase their effort by enough to raise their initial attempts by one letter grade (or 9.97 points to be exact) in hybrid classes to almost two letter grades (or 17.65 points) in fully online classes.

Conclusion

In this paper we study the impact of fail-forward (multiple attempt feedback) on student effort and learning in an introductory business statistics course. More specifically, students in various sections of business statistics courses were given three chances to complete quizzes during the semester. One treatment group was assigned the highest attempt score as the quiz score. The second treatment group was assigned the average score. We find that students score higher the first time around under the "average of all scores" grading schemes. Instructors who wish to use fail-forward techniques and need to ensure an adequate level of preparation of students are more

likely to obtain the desired outcomes by employing the more costly approach to grading (for the student), such as the "average of all scores" grading scheme used in this study.

While the "average of all scores" or high-penalty scheme seems to encourage effort, it may also be more stressful for students. Further research should look into student perceptions of the two grading schemes and their effects on student satisfaction and instructor evaluations.

Furthermore, even though the "average-of-all-scores" grading scheme seems to encourage effort, it would also be interesting to see if it ultimately improves student learning. A first indication can be found by looking at student scores in their last attempt at quizzes. Our data show that the highest-attempt-score was 6 points lower (on average) in online courses and 1 point lower (on average) in hybrid courses when the "average-of-all-scores" grading scheme was used. While this seems to suggest that students in fully online courses learn less under the "average-of-all-scores" grading scheme it could also mean that these students are content with fewer quiz attempts and slightly lower numeric grades (but the same letter grade). More research is needed to determine the ultimate effects of fail-forward grading schemes on student learning.

In this study the level of difficulty of quizzes remained the same in between attempts. Another way to encourage students to prepare well in advance of quizzes is to announce that the level of difficulty of questions can and will increase as the number of attempts increases. Further research is needed to determine whether or not this variation of the set up influences the level of preparation under both the "average of all scores" and the "highest of all scores" grading schemes.

In this study all feedback given to students was based on graded assignments. Future research should also look into whether or not effort and learning are affected by graded vs. ungraded feedback. Finally, we also need more research to understand whether or not feedback should be instantaneous or delayed, and if delayed, on the optimal amount of waiting-time.

References

Bangert-Drowns, R.L., Kulik, C.C., Kulik, J.A.., & Morgan, M. (1991). The Instructional Effect of Feedback in Test-Like Events. *Review of Educational Research*, 61 (2), 213-238.

Clariana, R.B. (1990). A Comparison of Answer until Correct Feedback and Knowledge of Correct Response Feedback under Two Conditions of Contextualization. *Journal of Computer Based Instructions*, 17 (4), 125-129.

Clariana, R.B. (1993). A Review of Multiple-Try Feedback in Traditional and Computer-Based Instruction. *Journal of Computer Based Instructions*, 20 (3), 67-74.

Clariana, R.B., & Smith, L.J. (1989). Comparative Research of Ability and Feedback Form in Computer-Based Instruction. *Paper presented at the Annual Meeting of the Mid-South Educational Research Association, Little Rock, AR.* Retrieved from http://www.personal.psu.edu/rbc4/ED436137.pdf. Retrieved on 10/20/2010

Coates, D. & Humphreys, B.R. (2001). Evaluation of Computer Assisted Instruction in Principles of Economics. *Educational Technology & Society,* 4 (2).

Heckman, James J., (1979). Sample Selection Bias as a Specification Error. *Econometrica: Journal of the Econometric Society*, 1979, *4 7*(1), pp. 153-62.

Kulhavy, R.W. (1977). Feedback in Written Instruction. *Review of Educational Research,* 47 (1), 211-232.

Patron, H. & W.J. Smith. (2009). Improving Students Performance in Business Statistics Courses with Online Assessments. *Journal of Economics and Finance Education*, 8 (2), 39-49.

Salomon, G., & T. Globerson. (1987). Skill may not be Enough: the Role of Mindfulness in Learning and Transfer. *International Journal of Educational Research,* 11, 623-637.

Winship, C., & R.D. Mare. (1992). Models for Sample Selection Bias. *Annual Review of Sociology*, 18, 327-350.