Incentives and Student Learning: A Natural Experiment with Economics Problem Sets

By Wayne A. Grove and Tim Wasserman*

Although economics professors carefully explain the importance and role of optimal decision making to their students, ironically they lack the most basic information regarding their own instructional production choices. The paucity of assessment studies of pedagogical practices results from the limited use of randomized experiments in higher education, apparently due to ethical concerns and institutional restrictions upon human subject research. Since sample selection problems always arise in the absence of a control group, the lack of experimental research design makes it difficult to confidently assess the efficacy of a particular instructional choice. Thus, little is known about what a faculty member can do differently to increase college student learning (William E. Becker, 1997; John J. Siegfried and William B. Walstad, 1998),[1] despite the fact that economics education research has been fostered since the 1950s by annual sessions of the American Economics Association and since 1969 with the *Journal of Economics Education.* Faculty-assigned problem sets, one of the most common teaching practices in college economics courses, are a case in point (Becker, 1997, Table 3, p. 1352). No study assesses the effectiveness of regular student practice of economic analysis, relative to a control group.

Fortuitously, then, in the fall of 1998, a natural experiment occurred in a principles of microeconomics course at Syracuse University, which divided 239 students into two groups, differentiated by whether or not assignments counted toward their course grade. In this paper, we estimate whether students with graded problem sets increased their mastery of course material relative to students without a grade incentive to do so, all else being equal. If grade incentives increased student learning, what was the magnitude of that treatment effect and which subgroups of students, if any, benefited most?

Our results indicate that the grade incentive to practice economics throughout the semester boosted the exam performance of one particular type of student: freshmen. Compared with control-group students who received identical assignments but no direct grade inducement to complete them, freshmen in the treatment group improved their Economics 101 exam performance by one-third to two-thirds of a letter grade, from a C to either a C+ or a B−. These findings about actions any professor can take to increase student learning should interest higher education stakeholders, especially those who wish to reverse high and rising college drop-out rates.

I. The Natural Experiment and the Data

During the fall of 1998, 239 Syracuse University undergraduate students enrolled in and completed four sections of introductory microeconomics taught by one of the authors at Syracuse University, a large, private residential university in the northeast (Carnegie Classification: Doctoral Research Universities II-Extensive). The average student was a 19-year-old white male. When a faculty member fell

[1] For exceptions, see John F. Chizmar and Anthony L. Ostrosky (1998) and Trisha L. N. Emerson and Beck A. Taylor (2004).

seriously ill as classes began, a natural experiment occurred that separated students into one group (of 143) whose course grades were based on problem set performance and another (of 96) whose course grades were not.[2] All lectures, handouts, exams, and review sessions were as similar as possible. Every student received the problem sets at the same time, as well as equal encouragement to practice economics by solving the assignments as important preparation for the exams. Students in both groups received the answer keys at the same time (when the graded-group students handed in their problem sets). Of the five assigned problem sets, two occurred prior to the first exam, two before the second exam, and one preceded the final exam. The problem set grade, the average of the best four assignment scores, comprised 15 percent of the final grade.[3] Two-thirds of the graded-group students completed all five problem sets, and only 10 percent handed in fewer than four assignments.

Exam scores and problem set data come from the author's records, while all other information used in this study come from university records (not from student surveys, which have been shown to overstate actual performance). Mastery of the course material was measured by performance on three required exams of equal value, each of which was assessed with a zero-to-100 point scale. The exams contained a combination of problem solving, short answer, and multiple choice questions. To provide as much uniformity in grading as possible, the same graduate student graded each question for all students. The order in which individual student exams and groups of exams were graded varied: teaching assistants either graded all the questions they were responsible for before passing exams on or marked one batch of exams at a time before swapping batches with the other graders.

[2] After the second week of classes, with the ill professor unable to return to the classroom and with the drop period passed, the newly assigned professor thought it inappropriate to impose new problem set requirements.

[3] Treatment group grades were comprised of exams (80 percent), problem sets (15 percent), and participation (5 percent), whereas control group grades consisted of only exams (95 percent) and participation (5 percent).

The data for this study come from a natural, rather than a perfectly randomized, experiment, since students enrolled in individual sections of Economics 101 without knowledge that a natural experiment would occur. We find no evidence of sample selection bias regarding either class size or the 18 students (out of the 96) in the nongraded group who had taken principles of macroeconomics the previous semester with the professor who fell ill (see Grove and Wasserman, 2005). The 27 students in each group who dropped the course were academically weaker, measured by that semester's GPA. With fewer nongraded group students, a higher fraction of them attrited (22 percent compared with 16 percent in the treatment group); thus, if anything, attrition biased down the treatment effect. With differences in right-hand-side variables controlled for, dropouts may have caused sample selection bias if some other unmeasured component of ability caused attrition. Finally, our results ameliorate sample selection concerns that students strategically enrolled in the section of Economics 101 with no problem sets in order to exert low effort (since freshmen had the least information about faculty and courses).

No SAT scores were available for 41 students, reducing the sample with complete data from 239 to 198 students (i.e., 20 from the graded and 21 from the nongraded group). To correct for these missing observations, a dummy variable was created and set equal to zero for missing SAT scores and equal to one for students with SAT scores. The estimated coefficient for each dummy variable represents the product of the average imputed value of the missing observation and its estimated coefficient. By retaining in the sample those records that have missing observations, this procedure makes it possible to use the information provided by the observations on the other variables in those records. Furthermore, the regression results are similar to and lead to almost identical conclusions as obtained when using only the sample of students whose records contained SAT scores.

Thus, the only discernible difference between the four sections of introductory microeconomics was that students in three sections had a direct grade incentive to practice economics problems throughout the semester (the treat-

ment group), whereas those in the control group received neither a direct reward for completing problem sets nor a penalty for failing to do so.

II. Grade Incentives and Learning

Of faculty who assign problem sets, some base a portion of the course grade on them and others do not. Why might graded assignments increase learning? Since ungraded assignments offer no direct grade benefit (or penalty), the indirect benefit of better exam or course performance must motivate students to complete them. The crux of students' cost-benefit calculus regarding problem set assignments is that the costs are certain and immediate (of foregoing some academic, athletic, social, or employment interest), but the payoff is uncertain and long run (e.g., end of the semester or of college).

A grade incentive might induce students to complete problem sets under three circumstances. First, if uncertainty is the obstacle ("will completing the assignments *really* increase my exam and course performance?"), graded assignments provide a direct reward/penalty (the magnitude of which depends upon its share of the course grade—for this study, 15 percent). Second, for those who think completing the homework will provide an indirect payoff (i.e., better course grade), but deem it too low to bother, graded assignments provide an additional benefit. Finally, for students who intend to, but do not, complete the ungraded assignments (i.e., those with time-management problems), a grade incentive/penalty may operate as a prioritizing mechanism, allowing such individuals to actualize their intentions. According to this line of reasoning, then, two sets of students would not change their behavior: those who would have completed the problem sets without any direct grade incentives and those for whom the expected payoff (direct and indirect) continued to remain too low, relative to the value of their time.

A variety of studies find that student aptitude influences how different teaching choices affect learning: practices that increase study discipline may benefit lower-GPA students whereas more time spent studying may benefit higher-aptitude students. Kris N. Kirby et al. (2005) found a negative relationship between discount rates and college GPA, suggesting that regular deadlines help below-average students.

III. Empirical Methodology

Motivating our work with an educational production function approach that is standard in the literature, we specify the following reduced-form model:

(1)   Student learning $= f$(academic ability;

demographic characteristics; membership

in a fraternity or sorority; experimental

group member).

We measure academic ability in two ways: collegiate GPA and math SAT scores (having shown elsewhere that they are the best scholastic control variables for performance in economics courses; Grove et al., 2006). Since GPA may measure comparative academic performance poorly due to the variability of grade distributions by faculty members, courses, and departments, we standardized semester GPA data by creating a *z*-score *for every class* that students in our sample took that semester (by subtracting raw grades in each course from the sample mean course grade and dividing that by the standard deviation of the course grades). Our measure of collegiate academic performance, zSemGPA-ECN, is students' standardized GPA excluding their Economics 101 grade the semester the experiment occurred.

Thus, hypothesizing that the incentive to practice economics throughout the semester improved college students' exam performance, we write our empirical model as:

(2)   *Mean Exam Score$_i$*

$$= \alpha + \beta(Graded\ Group)_i + \gamma X_i + \varepsilon_i$$

where X_i is a vector of observed student ability and demographic characteristics, namely standardized current semester college GPA without the economics grade, SAT math scores, whether the student's record contained an SAT

TABLE 1—ESTIMATED GRADED GROUP TREATMENT EFFECTS

Variable	(1)	(2)	(3)	(4)
Intercept	61.78**	61.61**	63.03**	62.72**
	(17.55)	(17.38)	(17.95)	(17.99)
zSemGPA-econ	7.37**	7.67**	7.30**	7.10**
	(13.43)	(9.03)	(13.46)	(11.80)
SAT-math	0.04**	0.04**	0.04**	0.04**
	(5.89)	(5.88)	(5.98)	(6.04)
Without SAT-math	16.83**	17.04**	16.52**	
	(4.60)	(4.61)	(4.57)	
Male	0.86	0.85	0.68	0.69
	(1.03)	(1.01)	(0.82)	(0.76)
White	−1.08	−1.13	−1.11	−0.89
	(−1.20)	(−1.25)	(−1.25)	(−0.92)
Freshman	−3.36**	−3.39**	−6.69*	−6.02**
	(−2.73)	(−2.74)	(−3.77)	(−3.15)
Sophomore	−3.47**	−3.50**	−3.86**	−3.98**
	(−3.04)	(−3.06)	(−3.39)	(−3.03)
Fraternity/sorority member	−1.98*	−1.97*	−2.10*	−2.60**
	(−2.11)	(−2.10)	(−2.26)	(−2.79)
Graded group	2.27**	2.25**	0.88	1.54
	(2.77)	(2.74)	(0.90)	(1.43)
Graded group × zSemGPA-econ		−0.51		
		(−0.47)		
Graded group × freshman			4.58*	3.74*
			(2.58)	(2.05)
Observations	239	239	239	198
Adjusted R^2	0.59	0.59	0.60	0.60

Notes: The dependent variable is the mean exam score in points on a 100-point scale; t-statistics are in parentheses.

   * Statistically significant at the 5-percent level.

  ** Statistically significant at the 1-percent level.

math score, race, gender, whether a freshman, whether a sophomore, and membership in a Greek organization.

IV. Results

We provide the results of four specifications of our estimated graded group treatment effect in Table 1, obtained using ordinary least squares. Conditioning on Greek organization participation, academic aptitude measures, and demographic characteristics, students with a grade incentive to study (the treatment) increased their average exam performance by an estimated 2.27 points on a 100-point scale, statistically significant at the 1-percent level (Table 1, column 1). This treatment effect would have boosted the average control group students' grade in Economics 101 by one-third of a letter grade, from a C+ (77.7) to a B− (80.0). Using the sample of 198 students whose records contained SAT scores (rather than the full sample, $N = 239$), we obtain a slightly larger treatment effect of 2.84 points in which the other coefficient estimates were essentially identical (not shown in Table 1; see Grove and Wasserman, 2005).

To test whether students of different academic aptitudes benefited differently from graded assignments, as suggested by existing studies, in specification 2 we include a term interacting membership in the treatment group with students' GPA. The coefficient on that interaction term is insignificant, quite small, and, most importantly, does not alter the magnitude or significance of the graded group variable; thus, we find no evidence that either academically above- or below-average students benefited more from graded problem sets, *ceteris paribus*.

To evaluate whether, and how, students with different demographic characteristics responded

to the grade incentive to practice, we compare, separately in succession, membership in the treatment group with each of the demographic variables, Greek system membership, and year in college. Only freshmen, according to our results, experienced a significant difference in performance due to the grade incentive to practice. Our estimates indicate that first-year students in the experimental group performed 4.58 points higher on exams (Table 1, column 3) relative to freshmen in the control group, controlling for aptitude and demographic variables. Since the average freshman in the control group had a 76.4 average exam grade, the 4.6-point treatment effect would have boosted their performance by two-thirds of a letter grade from a C to a B$-$ (81.0)—a very substantial improvement. While freshman scored 6.7 points worse than others on exams, first-year students with a grade incentive to complete the assignments made up two-thirds of their disadvantage of being new to college. Importantly, note that inclusion of the freshman-graded group interaction term caused the general treatment group effect to lose its significance, something that did not occur with any other combination of variables or interaction terms.

Unfortunately, because we have no information about nongraded group students' completion of assignments, we cannot evaluate whether freshmen in the treatment group relative to those in the control group, *ceteris paribus,* experienced larger gains in cognitive achievement due to completing more assignments. Among the 49 freshmen in the treatment group, assignment completion improved exam performance by 2.47 points per problem completed ($t = 2.38$).

As a check on the freshman treatment effect results, in specification 4, we replicate specification 3, using only the sample of students with SAT scores ($N = 198$). The coefficient on the freshman-graded group interaction term is 3.74 points, which implies a raise in the average control group freshman exam performance by one-third of a letter grade, from a C+ (77.1) to a B$-$ (81.7). The other estimated coefficients are essentially the same in specifications 3 and 4.

V. Conclusions

Using a natural experiment with a rich set of institutional and instructor data, we estimate

that a grade incentive to practice economics throughout the semester boosted the average freshman exam performance, but not that of academically above- or below-average students, or of any other category of students. The freshman treatment effect learning gain of 3.7 to 4.6 points increased first year students' mean grades from a C to either a C+ or a B$-$, *ceteris paribus,* relative to those in the control group. These statistically significant and meaningful results offer an exception to Eric Hanushek's (1991) suggestion that instructional variables are unimportant in explaining student learning and, most importantly, they identify one that can be implemented by the typical professor.

Although scholars often have found that freshmen perform worse in introductory economics courses, like so many of the empirical conclusions about student learning, those findings have led to no satisfactory policy solutions. An administrative ban on freshman in economics courses, for example, would cause difficulties for sequentially structured majors. By contrast, our results suggest a viable response to concerns about below-average freshman performance in economics courses and about student retention generally: provide professors with graders (or some other incentives) to encourage assignment of problem sets.

Since our data come from a single professor's classes in one semester, it will be important for other scholars to determine the generalizability of these findings. Given the obstacles and reluctance of academics to conduct randomized experiments, we hope faculty will recognize the importance of classroom natural experiments when they occur and use the data to test the efficacy of teaching methods and assignments.

REFERENCES

Becker, William E. "Teaching Economics to Undergraduates." *Journal of Economic Literature*, 1997, *35*(3), pp. 1347–73.

Chizmar, John F. and Ostrosky, Anthony L. "The One-Minute Paper: Some Empirical Findings." *Journal of Economic Education,* 1998, *29*(1), pp. 3–10.

Emerson, Tisha L. N. and Taylor, Beck A. "Comparing Student Achievement across

Experimental and Lecture-Oriented Sections of a Principles of Microeconomics Course." *Southern Economic Journal*, 2004, *70*(3), pp. 672–93.

Grove, Wayne and Wasserman, Tim. "Incentives and Student Learning." Unpublished Paper, 2005.

Grove, Wayne; Wasserman, Tim and Grodner, Andrew. "Choosing a Proxy Academic Aptitude." *Journal of Economic Education* (forthcoming).

Hanushek, Eric A. "When School Finance 'Reform' May Not Be Good Policy." *Harvard Journal of Legislation*, 1991, *28*(2), pp. 423–56.

Kirby, Kris N.; Winston, Gordon C. and Santiesteban, Mariana. "Impatience and Grades: Delay–Discount Rates Correlate Negatively with College GPA." *Learning and Individual Differences*, 2005, *15*(3), pp. 213–22.

Siegfried, John J. and Walstad, William B. "Research on Teaching College Economics," in William B. Walstad and Philip Saunders, eds., *Teaching undergraduate economics: A handbook for instructors*. New York: McGraw-Hill, 1998, pp. 141–66.