Southern Economic Journal 2004, 70(3), 672–693

Targeting Teaching

Comparing Student Achievement across Experimental and Lecture-Oriented Sections of a Principles of Microeconomics Course

Tisha L. N. Emerson* and Beck A. Taylor†

An increasingly popular alternative to the lecture-oriented "chalk-and-talk" approach to teaching principles of microeconomics is the use of classroom experiments. Like other alternatives to traditional teaching methods, there exists little more than anecdotal evidence supporting the effectiveness of the experimental approach. We estimate the effect of participating in classroom experiments on student achievement in a principles of microeconomics course. Nine sections (300 students) participated in the study, two of which (59 students) relied heavily on classroom experiments throughout the semester. The remaining seven sections (241 students) used no experiments. We find that students in the experimental sections experienced significantly higher gains in Test of Understanding in College Economics (TUCE) scores but differed little on other more qualitative outcomes. Additionally, results indicate that certain student characteristics, including gender, major, and grade point average, can be used to predict a student's likely success when choosing between courses that rely on experiments and those that employ more traditional forms of pedagogy.

1. Introduction

The 2002 Nobel Memorial Prize in Economic Sciences was awarded to Vernon Smith "for having established laboratory experiments as a tool in empirical economic analysis."[1] Professor Smith's contributions have had profound effects far beyond the theoretical and empirical advances of experimental economics for which he was honored. The use of experiments as pedagogical tools is also rapidly increasing. As evidence, the first principles of microeconomics textbook exclusively

* Department of Economics, Baylor University, Waco, TX 76798-8003, USA; E-mail Tisha_Nakao@baylor.edu.

† Department of Economics, Baylor University, Waco, TX 76798-8003, USA; E-mail Beck_Taylor@baylor.edu; corresponding author.

The authors gratefully acknowledge financial support from Baylor University. We also thank Susan Armstrong, Kate McGill, and the students and instructors who participated in our study at Baylor University for their patience and hard work during the data collection process. This paper has benefited greatly from the comments of Bill Becker, Stephanie Brewer, Eric Dearing, Steve Green, Chuck North, John Pisciotta, Mike Watts, two anonymous referees, and session participants at the 2002 Southern Economic Association conference. A portion of Taylor's work was completed while a visiting scholar at Harvard University. Any errors are the authors' sole responsibility.

[1] For more information, see http://www.nobel.se/economics/laureates/2002/press.html.

devoted to the use of experiments (Bergstrom and Miller 2000) is currently in its second edition and gaining wider acceptance.[2]

Rather than simply relying on the instructor-centered lecture that has historically been the preferred method of instruction to describe the market mechanism, the use of experiments in the principles classroom provides students with an experiential learning opportunity: the chance to participate in a controlled market environment and to observe market forces that are normally only talked about and described as movements on a graph. Experiential learning, followed by instructor-led discussion that places the experimental results into their theoretical contexts, affords the student with a hands-on learning experience that is quite different from the traditional "chalk-and-talk" approach. Bergstrom and Miller (2000, p. iv) summarize the student experience in the preface of their increasingly popular text:

> You will be studying the behavior and interactions of people in economic situations. And as one of these interacting economic agents, you will be able to experience first-hand the problems faced by such an agent. We suspect that you will learn nearly as much about economic principles from your experience as a participant as from your analysis as an observer.

While there is evidence that the experimental approach to teaching economics is being adopted by more instructors, there is relatively little empirical evidence regarding its efficacy as a pedagogical tool. Siegfried and Fels (1979) report initially that the use of games, simulation models, and demonstration routines in teaching economics is far from promising, with either no significant impact or a negative effect on student achievement. Further, early studies often focus on simply describing experiments, and these studies provide only anecdotal evidence as to their effectiveness and impact on student outcomes (see, e.g., DeYoung 1993; Williams and Walker 1993). On this note, Fels (1993, p. 365) comments that "it is ironic that those who use controlled experiments in their research on economics do not use controlled experiments to evaluate their teaching."

Recent studies take a more formal approach to estimating the value of experiments in the classroom. Gremmen and Potters (1997), for example, study the effectiveness on student learning of an international economic relations simulation game. They find that students who participated in the game performed better on objective tests than those students exposed to the material only through lecture. However, the Gremmen and Potters study, and others like it, focus only on the efficacy of a single experiment or a small set of experiments (see also Siegfried and Fels 1979; Frank 1997). Additionally, Fels (1993) questions the value of a single experiment in the context of comprehensive student achievement and suggests that, like a single classroom session, a single experiment is unlikely to make any important difference. Fels instead argues that studying the effectiveness of several experiments over the course of a semester is likely to provide a better indication of the value of experiments as a pedagogical tool.

Following this advice, two recent studies assess the effectiveness of multiple experiments in a principles of microeconomics course, but these studies find mixed results. Cardell et al. (1996) study

---

[2] For more information on the Bergstrom and Miller textbook, see http://www.econ.ucsb.edu/~tedb/eep/eep.html. Second edition sales of the Bergstrom and Miller text increased over the first edition by 5.6% and 48.7% in the U.S. and abroad, respectively. In addition to the Bergstrom and Miller text, many other resources are available to instructors who wish to use experiments in their classes (see, e.g., the Games Economists Play website developed by Delemeester and Brauer at http://www.marietta.edu/~delemeeg/games/ and Charles Holt's website at http://www.people.virginia.edu/~cah2k/teaching.html). Other available texts include instructor supplements containing sample experiments, and many of these experiments are made available online (see, e.g., Mankiw 2001). Additionally, Aplia, developed by Paul Romer, markets itself as a new generation publisher with online materials and experiments (see www.aplia.com).

a sample of 1800 students, with slightly less than one half assigned to experimental sections and the remainder assigned to traditional lecture-oriented sections. After controlling for a variety of student- and instructor-level characteristics, Cardell et al. find no statistically significant impact of the experimental approach on student achievement, where achievement is measured by differences in post- and precourse performance on the Test of Understanding in College Economics (TUCE). Dickie (2000), on the other hand, finds that, in his sample of 142 students, those exposed to experiments in their principles classes achieved a significantly greater improvement on the TUCE than those exposed to traditional pedagogy. Although students in the treatment group experienced greater TUCE improvement overall, Dickie also finds differential effects across students depending on their ability level. Specifically, Dickie finds that better students (as measured by grade point average [GPA]) experienced larger benefits from experiments while lower ability students may have experienced reduced achievement relative to what they may have attained under the traditional approach.

Becker (1997) notes that, although limited anecdotal evidence confirms the value of teaching with active approaches, these approaches have not been empirically demonstrated (at least not consistently) as superior to chalk-and-talk methods. Becker suggests that the failure in the literature to find consistent support for active teaching methodologies may lie in the testing methods employed rather than in the absence of any effect (e.g., inappropriate statistical methods and the use of assessment instruments with problematic measurement error). Further, Becker et al. (1991) call for the use of multiple measures of student outcomes as indicators of the efficacy of various teaching approaches, and they encourage the replication of earlier studies to investigate the robustness of previous findings.

The present study addresses the concerns and suggestions mentioned above and contributes to the small literature on the effectiveness of experimental methods by examining the potential differences in student achievement between students exposed to a comprehensive, multiexperiment approach to teaching principles of microeconomics and students participating in lecture-oriented classes that used no experiments. Our data, collected in the spring of 2002 at Baylor University, include a rich set of information on student achievement and other student- and section-level characteristics from a sample of 300 students. Specifically, two out of nine total sections of the core microeconomics course (59 students) used 11 experiments from the Bergstrom and Miller (2000) textbook to supplement the curriculum. The remaining seven sections (241 students) used the traditional lecture-oriented approach. After controlling for student- and section-level characteristics, our results indicate that students in the experimental sections improved their TUCE scores by an average of 2.42–2.99 points over the control group, or using a slightly different measure, by 11.1–12.3 percentage points of the possible percent increase in scores. Additionally, these differences are present across various cognitive, content, and difficulty levels. We find few differences between experimental and non-experimental students, however, in other outcomes, such as performance on a departmental final exam, student evaluations, or class attrition rates. Finally, we find that certain student characteristics can affect the likelihood of achievement in an experimental course, including a student's gender, GPA, and major. These results are robust to potential issues of positive selection bias, endogeneity of precourse ability, and potential censoring of the dependent variable.

The remainder of this article is organized as follows. Section 2 describes the data and presents our empirical methodology. Section 3 discusses the empirical results on various qualitative and quantitative student outcomes and presents differential effects across student characteristics that can provide a road map for academic advisors or course coordinators wishing to advise students as to their likely benefit from participating in an experimental section. Finally, section 4 summarizes our results and discusses possible extensions of this research.

2. Data and Empirical Methodology

Students in our study were enrolled in one of nine sections of the core course in microeconomics principles at Baylor University during the 2002 spring semester.[3] Two of these sections (the treatment, or experimental, group consisting of 59 students) supplemented the standard curriculum using 11 in-class experiments taken from the Bergstrom and Miller (2000) textbook. The remaining seven sections (the control group consisting of 241 students) used the traditional lecture-oriented methodology.

Aside from the treatment group's use of experiments, considerable effort was made to maintain as much homogeneity as possible, both between and within the control and treatment groups. The number of total contact hours between students and instructors was equal across all sections, though the allocation of these hours differed across control and experimental sections, with experimental sections substituting experiments for lecture time and other activities.[4] All students in the sample used the same required textbook (a commonly used microeconomics principles text) and covered the same major topics.[5] Assignments across all sections were similar, including a mixture of homework, two or three midterm exams, and a comprehensive final exam. Each section of the course employed exams that consisted of some combination of multiple-choice and essay questions. Finally, the class size of all sections fell within the range of 23–35 students.

Both sections within the experimental group were organized in the same manner. Students in this group participated in one experiment per week (usually taking one full class period), while the remaining class time was devoted to lecturing on theoretical concepts and reconciling those concepts to the data generated from the experiments. Experimental students were assigned a homework assignment each week that was based largely on the experiment and the interpretation of the experimental outcomes. Table 1 presents a description of the 11 experiments used.

Model of Student Learning

To motivate our empirical work, we use an educational production function approach that is standard in the literature (see, e.g., Siegfried and Fels 1979). In this approach, the following reduced-form model is specified:

> Student learning = *f*(aptitude; educational background; other student-specific characteristics; educational environment, technology, or teaching methodology; observed and unobserved instructor- and section-level effects).

Our student learning measure takes two forms: (1) the absolute difference between post- and precourse scores on the Test of Understanding in College Economics (TUCE), and (2) a gap-closing measure defined as the difference in post- and precourse TUCE scores expressed as a percentage of the maximum possible point improvement available based on the student's precourse TUCE score.[6] In

---

[3] Although students were not randomly assigned, they did not know *ex ante* whether they had enrolled in an experimental section.

[4] For example, lecture accounted for an average of 45% of total contact time in the two experimental sections compared with an average of 55% of total contact time for control sections. In general, control sections relied more heavily on other activities such as class discussions and presentations.

[5] The curriculum was standardized so that instructors across all sections were required to present the same major topics to their students. We recognize, however, that the amount of time spent on each topic could have varied across instructors, but an informal poll of the instructors in our sample indicates that this was not the case.

[6] In other words, the gap-closing measure is defined as (postcourse TUCE – precourse TUCE)/(33 – precourse TUCE). We also examined two other dependent variables: the postcourse TUCE score alone and the unweighted percentage change in post- and precourse TUCE scores. Results are qualitatively unchanged when using these other measures.

Table 1. Bergstrom and Miller Experiments Utilized in Experimental Sections

Experiment Name	Brief Description
1. Supply and demand	Buyers and sellers are given reservation values for a homogeneous product. Trades are carried out in a double oral auction environment. Successive rounds yield prices close to the equilibrium prediction. This experiment serves as the model for many of the future experiments.
2. Shifting supply	Comparative statics are analyzed as the market experiences exogenously imposed supply shocks. Students observe that prices converge near the new equilibrium prediction.
3. Minimum wages	An artificial price floor is imposed in a labor market. Students observe the surpluses generated by the price control.
4. Sales tax	Tax incidence is examined via a sales tax imposed on the market. Students confirm the allocative equivalence of taxes imposed on either sellers or buyers.
5. Externalities	Students' surpluses are reduced to reflect an external cost of production. Students learn the effectiveness of using Pigouvian taxes or a pollution permit auction to internalize the external cost and correct the market failure.
6. Measuring productivity	Students participate in production teams using fixed and variable inputs. Diminishing marginal productivity is demonstrated and corresponding cost concepts are discussed.
7. Entry and exit	Long-run competitive equilibrium is obtained as market participants exit and enter the market depending on their profitability. The distinction between economic and accounting profit is emphasized. The role of sunk costs is discussed.
8. Monopolies and cartels	Firms participate in cartels and are forced to make joint pricing decisions. This restriction is relaxed and reversion to the competitive equilibrium is observed as defection occurs. Emphasis is placed on marginal revenue and marginal cost calculations.
9. Network externalities	Collective actions are highlighted as students must decide whether to participate in several new-technology markets. Buyer values depend on the number of other buyers in the same market. Network externalities, multiple equilibria, and stability concepts are demonstrated.
10. Comparative advantage	Production possibilities are imposed on two economies. Students learn that specialization and trade allow both economies to consume beyond production possibilities. Emphasis is placed on the difference between absolute and comparative advantage.
11. Adverse selection	A "market for lemons" is created using buyer/seller information asymmetry. Students observe a shortage of high-quality products in the market. Signaling and screening concepts are covered.

addition to measures of aptitude (e.g., students' GPA or Scholastic Aptitude Test [SAT] scores), educational background (e.g., a student's major and whether a student has taken high-school economics) and other student-specific characteristics (e.g., gender and ethnicity), we include a dummy variable for the treatment group (experimental section) that captures the differential effect, if any, on student learning associated with the experimental treatment. We discuss our controls for instructor- and section-level effects in more detail later in this section.

Measure of Student Learning (TUCE)

To measure student learning, we administered the 33-question microeconomics portion of the TUCE to all students in the sample on the first and last days of class. This research design allows us to measure differences in learning, or value-added, across students. To provide incentive to exert effort on the precourse TUCE, students were informed that their performance on the precourse TUCE would impact their final course grade, but were not told explicitly how their score would be included in this calculation.[7] To induce effort on the postcourse TUCE, students were informed that their course grades would be based, in part, on their improvement over their precourse TUCE score.[8] To further preserve the integrity of the TUCE data, no instructor in either the treatment or control groups was given access to the TUCE (either pre- or postcourse) and, thus, no inadvertent teaching to the test was possible (Gramlich and Greenlee 1993).

It is important to note that Becker (1997) and others question the validity and reliability of the TUCE as a measure of student learning. Potential problems with the TUCE include questions regarding the adequacy (or lack thereof) of an objective, fixed response test for measuring learning, and a concern over a student's ability to apply their limited knowledge to the real-world issues that appear in the TUCE. Becker (1997, 2001) also calls for assessment of other student outcomes (e.g., number of majors, attrition rates, student's perception of achievement, etc.) that are also of importance when assessing different pedagogical methods. We acknowledge that there are likely limitations in using the TUCE and that many of these limitations stem from the inherent measurement error associated with any testing instrument. However, the TUCE is a nationally recognized measure of student achievement in economics, many recent studies like ours use the TUCE, and few good substitutes are available if one is interested in a quantitative measure of achievement. One common substitute for the TUCE found in the literature is the use of final course letter grades. It is our opinion, however, that problems with this measure (e.g., subjectivity and comparability across instructors) are even more severe than many of the issues involving the TUCE. Additionally, if one is interested in measuring the value-added from course content or a particular pedagogical method, a measure of change is needed, and course grades cannot serve this purpose. Finally, the TUCE continues to be used extensively in the economics education literature, and our use of the TUCE affords greater consistency and comparability with this literature.

Importantly, however, we fully appreciate the value of measuring other student outcomes. To

---

[7] The precourse TUCE was designed to be a surprise exam; that is, students were to have had no knowledge of the exam before coming to class because such knowledge could have affected attendance and participation in the study. Because some sections began on Monday and others on Tuesday, there is some possibility that some students taking the exam on Tuesday knew of the exam beforehand. Thus, models (described later) that explicitly account for possible selection bias include as a control whether the class met first on Monday or Tuesday. In general, this day-of-the-week effect is not present in either the selection issue or other empirical findings presented herein.

[8] During the precourse TUCE assessment, students were not made aware of this grading method to prevent strategic behavior that could have led to a downward bias of the precourse TUCE scores. Additionally, care was taken to ensure that the change in TUCE scores entered into grade calculation in a similar fashion across all sections.

this end, we collected additional outcomes across the control and treatment groups. These measures include student performance on a set of 17 common multiple-choice questions from the departmental final exam, student evaluations of the course and instructor, the student's self-reported likelihood of taking future economics courses, student absences, and attrition rates.

Measuring Other Inputs in the Educational Production Function

The production function approach described above models student learning as a function of aptitude, educational background, other student-specific characteristics, the educational environment, and instructor- and section-level effects. In addition to TUCE scores, we collected data on students' cumulative grade point average at the beginning of the course (GPA, 4-point scale), math and verbal SAT scores, current and previously completed semester hours, the number of previous attempts in the same microeconomics principles course, average weekly work hours from employment, whether the student had taken economics in high school, the total number of student absences, and students' major, gender, and ethnicity.[9]

Maxwell and Lopus (1994) demonstrate that students' self-reporting of GPA, SAT scores, and other variables of interest may suffer from systematic reporting error. Such reporting error could potentially produce biased estimates of the relationship between student achievement and educational inputs. Thus, as in Chizmar and Ostrosky (1998), we collected GPA, SAT, current and previously completed semester hours, and the number of previous attempts in the microeconomic principles course directly from students' official university records. Student absences were reported by instructors. The remaining data were collected from student surveys.[10]

Summary statistics (unconditional means and standard deviations) for the pre- and postcourse TUCE scores as well as the explanatory variables used in our analysis, across control and treatment groups, are presented in Table 2. Students in the experimental and control sections scored similarly on the precourse TUCE, but as Table 2 reports, students participating in the experimental sections achieved significantly higher scores on the postcourse TUCE, a difference of approximately 2.17 points. Students in the control and experimental groups were similar in most other attributes of interest. With the exception of the number of semester hours completed and the number of student absences, there were no significant differences in the means between the two groups. Students in the experimental group had completed fewer semester hours at the beginning of the course and were absent more often than students in the control group. Both of these differences would tend to reduce any expected improvement in the treatment group's TUCE score, other things equal, as they indicate a lower level of educational experience prior to the course and less exposure to the course material, respectively (Durden and Ellis 1995).[11]

---

[9] We examine both standardized and unstandardized student absences. In the calculation of standardized absences, absences in a 2-day per week schedule (meeting Tuesdays and Thursdays) were multiplied by 1.5 to standardize absences to a 3-day per week schedule (meeting Mondays, Wednesdays, and Fridays).

[10] Two surveys were administered, one at the beginning and another at the end of the course, each prior the administration of the TUCE. Generally, the precourse survey asked for responses on student-level characteristics, like gender and ethnicity, whereas the postcourse survey asked for responses concerning instructor evaluations, likelihood of taking additional economics classes, and other questions about the students' perceptions of the quality of the course.

[11] Also of potential interest are class size differences (Raimondo, Esposito, and Gershenberg 1990; Kennedy and Siegfried 1997; Becker and Powers 2001). The mean class size in the treatment group is 28.2 versus 33.5 in the control group. This difference is statistically significant at the 5% (two-tailed) level. Collinearity between class size and the experimental dummy variable ($r = -0.64$) prevents us from controlling for both effects simultaneously. However, when significant class size effects are present in the existing literature, they are often a result of much larger class size differentials than we observe in our data (e.g., between small classes of 25–35 and large classes of 200–350 students).

Table 2. Descriptive Statistics (Experimental vs. Nonexperimental Sections)

Variable	Nonexperimental Mean (SD)	Experimental Mean (SD)
Precourse TUCE	9.89	9.44
	(3.31)	(2.76)
Postcourse TUCE	13.94	16.11[a]
	(4.62)	(4.49)
GPA	2.87	2.75
	(0.69)	(0.71)
SAT math	573.55	575.54
	(80.78)	(78.97)
SAT verbal	548.53	544.29
	(76.15)	(64.27)
Male	0.60	0.61
	(0.49)	(0.49)
Nonwhite	0.12	0.17
	(0.33)	(0.38)
Number of previous attempts	0.09	0.15
	(0.28)	(0.41)
Work hours per week	6.27	4.13
	(9.80)	(8.59)
Semester hours completed	43.34	34.63[a]
	(20.10)	(19.18)
Current semester hours	14.21	14.07
	(1.88)	(1.59)
Number of absences	2.70	4.62[a]
	(2.66)	(4.45)
Standardized number of absences	3.28	5.42[a]
	(3.12)	(5.80)
High-school course in economics	0.78	0.86
	(0.41)	(0.35)
Business student	0.74	0.76
	(0.44)	(0.43)
Number of observations	241	59

[a] Experimental and nonexperimental means are statistically different at the 5% (two-tailed) significance level or better.

Notes on Estimation Methods

By their very nature, both of our dependent variables (a change measure defined as the difference in post- and precourse TUCE scores, and this change as a percentage of possible improvement) are potentially subject to censoring problems (Siegfried and Fels 1979). In our data, for instance, the simple change measure has a potential range between the values of –33 and 33, and values of the gap-closing measure potentially range from $-\infty$ to 1. In our sample, however, the change measure actually ranges between –6 and 20, and the gap-closing measure falls between –0.30 and 0.71. Given the potential censoring issue, we estimated all of our models using a Tobit estimation procedure. Our results were robust to a variety of specifications that differed in their assignment of mass points in the distributions of our dependent variables.

Another potential problem in studies like ours is positive selection bias. It is conceivable that the same factors influencing a student's performance on our dependent variables may also influence whether a student persists in the course long enough to take the postcourse TUCE or whether the student was

present to take the precourse TUCE. If pre- or postcourse scores are missing in a systematic manner, failure to control for sample selection will result in biased estimates, where positive selection bias is often speculated to be the most likely result. Becker and Powers (2001) advocate the use of a standard Heckman selection correction to control for any potential nonrandom attrition from the sample. In our sample of 300 students, 37 values of the dependent variable are missing because we have only the postcourse TUCE scores for 11 students (i.e., students who joined the course after the first day of class), and we have only the precourse TUCE scores for 26 students (i.e., students who took the precourse TUCE but either withdrew from the class or missed the examination). To control for any potential selection bias, we employ the Heckman approach and, not surprisingly, find statistically significant selection issues in our data. Consistent with previous literature, GPA and SAT scores are each negatively and significantly correlated with the incidence of missing data, and the number of student absences is positively and significantly correlated with missing data.[12] Importantly, however, estimates of the coefficients on the predictors of student learning, particularly our treatment-group dummy variable, are materially unchanged when controlling for selection.

Becker and Salemi (1977) convincingly demonstrate that the inclusion of the precourse TUCE score as an explanatory variable in an education production function model like ours may introduce endogeneity and result in biased and inconsistent coefficient estimates. The precourse TUCE score is generally considered a proxy for precourse aptitude, either general or specific to economics, and is included as a control when we use the change in TUCE scores as our dependent variable. Because the precourse score measures this aptitude with some error, its use as a proxy for aptitude introduces bias into an ordinary least squares (OLS) estimation via its correlation with the OLS error term. As such, Becker and Salemi suggest using a two-stage least squares approach in which instruments are used for the precourse TUCE score. We estimated our model using this instrumental variables approach using each student's GPA, SAT scores, and whether the student had taken high-school economics as instruments for the precourse score. While our instruments are significant and positive predictors of the precourse score (each $p < 0.01$) and our first-stage model fit is good (average adjusted $R^2 = 0.39$), using the two-stage least squares procedure does not materially affect the results on our variable of interest—the effect of being in the treatment group. Additionally, the qualitative effects of our other explanatory variables, while generally measured with less precision with the instrumental variables approach, are also unchanged.[13]

Thus, while censoring, selection, and endogeneity of precourse scores are each a potential issue in our data, we find no qualitative (or significant quantitative) differences in our estimates of the effects of participation in the control group when we specifically account for these issues. Thus, in the results that we present below, we report only OLS estimates.[14]

An additional concern related specifically to our experimental design is the difficulty we have in measuring unobserved instructor-level effects. Although one preferred research design would include having each instructor teach both control and treatment sections, resource constraints prevented us from employing this design.[15] Instead, two instructors in our study taught the two sections in the

---

[12] Instructor dummy variables and the day of the week the precourse TUCE was taken (i.e., Monday or Tuesday) were used to identify the selection equation. Note that the day of the week the precourse TUCE was taken was statistically insignificant in the selection equation.

[13] Using two-stage least squares, the coefficient on the precourse TUCE score became positive and statistically significant, consistent with the results of Becker and Salemi (1977).

[14] The estimates from the Tobit, Heckman, and two-stage least squares procedures are available from the authors on request.

[15] Random assignment of instructors to sections is also a potential solution. In this scenario, sections would be listed in the course catalog without information on the instructor or teaching method. Instructors would then be randomly assigned to each section. This method would also pose significant constraints on resources and is likely to be practically infeasible.

treatment group and the remaining five instructors taught the other seven sections in the control group. As such, our treatment control variable is perfectly collinear with the instructor dummy variables, and we are unable to directly estimate via fixed effects any unobserved instructor characteristics separately from the experimental effect. This becomes a problem (i.e., a biased estimate of the treatment effect) if, for any reason, the two instructors teaching the experimental sections had characteristics other than using the experimental methodology that were also correlated with student performance. However, it is likely the case that those unobserved instructor characteristics most likely to affect students' performance on the TUCE exam are also correlated with observed section-level differences, namely the distributions on other student performance measures included in our data, specifically performance on the departmental final exam and student evaluations of the instructor and course. Thus, while we cannot include instructor-level dummy variables in the estimation, we can include section-level information on the distributions (i.e., means and standard deviations) of the final exam and student evaluations, measures that are likely influenced by unobserved instructor- and section-level differences. Though not ideal, this technique will likely capture much of the variance in student performance accounted for by the variance in unobserved instructor heterogeneity. Further, there is some empirical evidence that this approach is reasonable. Fleisher, Hashimoto, and Weinberg (2002) find little difference across 68 graduate teaching assistants after controlling for student characteristics. Watts and Bosshardt (1991) find a wide variance in instructor effects, but find that this variance is inversely related to the degree of coordination across instructors. Because we purposely imposed a significant level of coordination across instructors, both within and between the control and treatment groups, Watts and Bosshardt's results suggest a lower level of potential variance across instructors in our study. Further, Shmanske (1988) finds no significant difference across 17 principles instructors at one university.

Finally, we recognize that the errors across students in the same section (with the same instructor and subject to the same environmental and peer effects) are likely to be correlated. That is, our empirical model may be written as

$$\text{student learning}_{ij} = \alpha + X_{ij}\beta + \varepsilon_{ij}, \text{ where } \varepsilon_{ij} = \nu_j + \xi_{ij}.$$

In other words, learning of student i in section j is a function of a constant, student-specific characteristics (X_{ij}, including aptitude, educational background, gender, etc.), and an error term (ε_{ij}). This error term can be broken into two separate components, one component common to all students in section j (ν_j) that is dependent within section but independent across sections, and an idiosyncratic student $\times$ section error term (ξ_{ij}) that is independently and identically distributed across all students. By controlling for this potential interdependence in the error term for students in the same section, we also likely account for some portion of any unobserved instructor-level heterogeneity present in our data that would bias estimates of standard errors.

3. Results

Qualitative Outcomes

As we discussed earlier, the performance of students on an objective test is but one of a variety of outcomes of interest when evaluating different teaching approaches (Becker 1997, 2001). Before we present results of the regression analyses using TUCE scores, we first report a set of additional outcomes for our control and experimental groups in Table 3.

Table 3. Various Outcomes (Experimental vs. Nonexperimental Sections)

Outcomes	Nonexperimental Mean (SD)	Experimental Mean (SD)
Number correct on 17 common final-exam questions	13.77	13.66
	(2.07)	(1.89)
Student evaluations: (strongly agree = 5; agree = 4; neither agree nor disagree = 3; slightly disagree = 2; strongly disagree = 1)		
Instructor appeared interested in the subject material	4.64	4.77
	(0.70)	(0.42)
Instructor stimulated my interest in this subject	4.00	3.89
	(0.94)	(1.15)
Instructor stimulated my thinking	4.08	4.09
	(0.93)	(0.95)
Assignments contributed to my understanding of the course content	4.13	3.81[a]
	(0.82)	(0.98)
I learned a great deal from this course	4.13	4.04
	(0.90)	(0.96)
Student's expected course grade (A = 4.0; B = 3.0; C = 2.0; D = 1.0; F = 0.0)	3.05	2.87
	(0.76)	(0.79)
Student's actual course grade (A = 4.0; B = 3.0; C = 2.0; D = 1.0; F = 0.0)	2.65	2.27
	(1.01)	(1.14)[a]
Student's likelihood of taking future economics courses (none = 0; slight = 1; average = 2; high = 3; very high = 4)	2.97	3.02
	(0.26)	(0.31)
Student agreed that academic background prepared him/her for the class	0.61	0.53
	(0.49)	(0.50)
Hours per week that the student spent on preparation for class	3.74	3.92
	(2.17)	(2.36)
Student's number of absences during the semester	2.70	4.62[a]
	(2.66)	(4.45)
Student's number of standardized absences during the semester	3.28	5.42[a]
	(3.12)	(5.80)
Student dropped course during the semester	0.06	0.10
	(0.24)	(0.30)
Student's ranking of the five most valuable course activities (from most selected to least selected)	Homework Readings Lectures Quizzes	Experiments Lectures Readings Homework
Student's ranking of the five least valuable course activities (from most selected to least selected)	Lectures Readings Homework Projects	Homework Readings Experiments Lecture
Number of observations	241	59

[a] Experimental and nonexperimental means are statistically different at the 5% (two-tailed) significance level or better.

A portion of the comprehensive final exam administered to all students in each of the nine sections in our study consisted of a common set of 17 multiple-choice questions covering the major topics in the course. These questions were written by the seven instructors participating in our study, with each instructor responsible for two or three questions. The purpose of the common final-exam questions was to differentiate between average and below-average student understanding and mastery of the material (i.e., to differentiate between letter grades of C and below), as opposed to assessing

advanced levels of achievement. As Table 3 reports, the mean scores of the control and experimental groups on this common portion of the final exam are not significantly different. The failure to find a difference in the performance on the final exam between the two groups is not surprising given the basic nature of the questions asked. Additionally, this set of common multiple-choice questions is not a validated instrument and is susceptible to potential bias (i.e., the majority of the instructors writing the exam taught the control sections and were thus more likely to write questions better suited for the traditional lecture-oriented approach). Further, given the availability of the exam to all of the instructors, inadvertent teaching to the test cannot be ruled out. Importantly, however, mean final-exam scores did differ significantly across sections. This is important to our use of section means on the final exam to control for any unobserved instructor heterogeneity in the regression analysis to follow.

Table 3 also reports summary statistics on students' evaluations of the instructor and course. With the exception of the level of student agreement to the statement "assignments contributed to my understanding of the course content," the responses of the treatment group are not significantly different from those of the control group. Again, it is important to note, however, that the cumulative evaluation scores (added together) did differ significantly across sections, making this cumulative evaluation score, along with section-level information on the final exam, a potential instrument for unobserved instructor heterogeneity. Concerning the difference in responses on the one evaluation measure that is significantly different, the mean response of students in the experimental group is 3.81 compared with 4.13 in the control group. While this difference is statistically significant, a difference of 0.32 on a 5-point scale (5 = strongly agree to 1 = strongly disagree) is questionable in its practical significance. A failure to find a statistically significant difference in the responses of the two groups in their level of agreement with the statement "I learned a great deal from this course," also deserves specific mention. Although students from the control and experimental groups do not perceive a significant difference in their learning, we find that, on average, students in the experimental group had an unconditional increase in their score on the TUCE by more than 2 questions over the control group. These results are not surprising in light of the literature. Gremmen and Potters (1997) find no significant relationship between what students think they learn and what they actually learn, and Gramlich and Greenlee (1993) find relatively little correlation between student performance and student evaluations of teaching performance. However, we cannot discount the result that sample means of the students' evaluations in our data do not support generally the idea that students exposed to the experimental method of teaching have a higher level of satisfaction with the course at the end of the semester.

Using surveys, students were also asked to rank the activities (e.g., lectures, homework, experiments, exams/quizzes, etc.) they found most and least valuable during the course. Students in the treatment group ranked experiments as the most valuable course activity, with lectures second, and students in the control group ranked lectures third after homework and quizzes. We believe that this points to the complementarity (and not substitutability) between experiments and lectures. Learning theory indicates that students learn best when a concept is presented from a variety of angles, and that repetition of ideas in a range of contexts is vital for student learning (Fels 1993). Experiments in concert with lectures and discussion provide just such an approach.

Also reported in Table 3 are data on attrition rates and student absences. The attrition rates are not significantly different between the control and treatment groups. Regarding student absences, we find that students in the treatment group were absent more often. This outcome is particularly interesting in light of our main finding that the treatment group outperformed the control group. That is, despite attending fewer classes, the treatment group was able to learn more as measured by differences in post- and precourse TUCE scores. Becker (1982) presents a constrained optimization

model of student behavior and demonstrates that improved pedagogical methods need not necessarily lead to greater learning. In Becker's model, students are able to more efficiently transform effort into learning through these improved techniques, choose the same overall level of learning, and then reallocate their extra time to other utility-maximizing activities. In this context, our results concerning absences may indicate that experiments are a more efficient method than the traditional lecture-oriented approach.[16]

Additionally, Becker (1997, 2001) calls for studies on the effects of pedagogical tools on subsequent student enrollment in economics courses and selection of major. While our study is too recent to afford us a sufficient longitudinal dataset for this purpose, we report students' self-reported likelihood of taking future economics courses. We find no significant difference between the two groups on this point. Factors other than pedagogical approach, however, are likely to affect student responses to this question. Three fourths of our sample are business majors who are required to take both microeconomic and macroeconomic principles, with the macroeconomics course generally the second course taken in the sequence. The effect on enrollments in post–principles classes would serve as a better indicator of any differential pedagogical effect on student course selection.

Regression Analysis of Student Learning

We estimate the impact of the experimental pedagogical approach on our measures of student learning for our usable sample of 263 students.[17] Tables 4 and 5 report estimates for our change measure and gap-closing measure, respectively. As we discuss in section 2, all estimates are obtained using OLS, but these estimates are robust to corrections for sample selection, censoring, and potential endogeneity.

Table 4 presents six specifications of the change model. The first specification is a simple difference-in-means estimation. Students in the lecture-oriented sections experienced, on average, an improvement of 3.9 questions while students in the experimental group gained an average of 6.9 questions, an additional 3 questions above the control group.[18] Specification (2) adds a control for the precourse TUCE score. With this addition, the differential effect of the experimental treatment decreases slightly to 2.7 questions.

Specifications (3) and (5) include additional controls for student-specific characteristics, including aptitude, educational background, major, gender, and ethnicity. With respect to our aptitude measure, we find a high level of correlation between GPA and SAT scores ($r = 0.69$) and therefore use them alternately as proxies for aptitude in Specifications (3) and (5), respectively. Kennedy and Siegfried (1997) find that SAT scores serve as a better predictor of postcourse TUCE scores than does GPA. We report both specifications because the measures are associated with different sample sizes (263 for GPA, 235 for SAT).[19] We find that both measures of aptitude are positively correlated with student learning, which has been an extremely robust finding in the literature (Becker 1997).

---

[16] Of course, one should not interpret our statements as advocating less frequent attendance. To the contrary, empirical results presented below indicate that, other things equal, student learning suffers as a result of missing class, especially for those students in an experimental section.

[17] Recall that 37 students had missing values for either the pre- or postcourse TUCE score.

[18] The U.S. national average improvement is 4.32 questions on the 33-question microeconomics portion of the TUCE.

[19] Twenty-eight students of our usable sample did not have SAT scores either because they were transfer students who were not required to report their SAT scores to the university or because they substituted the American College Test (ACT) for the SAT. Kennedy and Siegfried (1997) find that ACT scores are a relatively poor measure of ability to learn economics; thus, we opted not to use standard conversion tables available for converting ACT to SAT scores, using instead only those 235 observations for which we have actual SAT scores.

Table 4. Absolute Difference in Post- and Precourse TUCE Scores

	Specifications					
Independent Variables	(1)	(2)	(3)	(4)	(5)	(6)
Intercept	3.893***	7.595***	2.181	−18.074	−7.489**	−25.279
	(0.399)	(0.828)	(2.424)	(18.506)	(2.300)	(20.928)
Experimental section	2.985**	2.709**	2.568**	2.592**	2.529**	2.416**
	(0.960)	(0.882)	(1.031)	(0.989)	(1.004)	(0.938)
Precourse TUCE		−0.371***	−0.475***	−0.489***	−0.674***	−0.677***
		(0.073)	(0.054)	(0.051)	(0.067)	(0.065)
GPA			1.618**	1.579**		
			(0.517)	(0.498)		
SAT math					0.011***	0.011***
					(0.003)	(0.003)
SAT verbal					0.021***	0.020***
					(0.003)	(0.003)
Male			1.537***	1.598***	0.933*	0.969*
			(0.439)	(0.426)	(0.426)	(0.441)
Nonwhite			−0.864	−1.056	−0.985	−1.168
			(0.713)	(0.799)	(0.933)	(1.032)
Number of previous attempts			2.306*	2.284*	1.835	1.779
			(1.211)	(1.204)	(1.425)	(1.371)
Work hours per week			−0.065	−0.064*	−0.059*	−0.058*
			(0.035)	(0.033)	(0.030)	(0.030)
Semester hours completed			0.004	0.003	0.014	0.012
			(0.029)	(0.030)	(0.024)	(0.024)
Current semester hours			0.035	0.046	−0.040	−0.029
			(0.151)	(0.144)	(0.166)	(0.164)
Number of standardized absences			−0.027	−0.013	−0.121	−0.111
			(0.078)	(0.074)	(0.100)	(0.107)
High-school course in economics			1.152*	1.166*	0.900	0.870
			(0.515)	(0.555)	(0.598)	(0.647)
Business student			−0.375	−0.301	0.090	0.180
			(0.561)	(0.592)	(0.653)	(0.685)
Final exam section mean				1.394*		1.421
				(0.671)		(1.284)
Final exam section SD				5.951*		6.988*
				(3.088)		(3.101)
Student evaluation section mean				−0.325*		−0.450**
				(0.165)		(0.151)
Student evaluation section SD				−0.745		−1.204*
				(0.567)		(0.626)
Observations	263	263	263	263	235	235
Adjusted R^2	0.06	0.13	0.20	0.23	0.34	0.36

Robust standard errors are in parentheses and are adjusted for within-section correlation of errors.
* Significant at 10%; ** significant at 5%; *** significant at 1%.

We again see in Specifications (3) and (5) that the experimental students have significantly larger point gains on the TUCE exam, other things equal. The remaining coefficient estimates, when statistically significant, have the expected sign. Males and whites tend to perform better, though ethnicity is never statistically significant. Previous exposure to economic concepts through either previous attempts at taking the course or having studied economics in high school improves student

Table 5. Gap-Closing Measure (Postcourse Minus Precourse TUCE/33 Minus Precourse TUCE)

	Specifications				
Independent Variables	(1)	(2)	(3)	(4)	(5)
Intercept	0.166***	−0.048	−0.924	−0.379**	−1.144
	(0.017)	(0.116)	(0.845)	(0.135)	(0.963)
Experimental section	0.121***	0.120**	0.115**	0.123**	0.111**
	(0.036)	(0.045)	(0.044)	(0.047)	(0.045)
GPA		0.061**	0.059**		
		(0.023)	(0.023)		
SAT math				0.000**	0.000**
				(0.000)	(0.000)
SAT verbal				0.001**	0.001**
				(0.000)	(0.000)
Male		0.063**	0.064**	0.038**	0.038**
		(0.019)	(0.019)	(0.015)	(0.015)
Nonwhite		−0.043	−0.050	−0.048	−0.055
		(0.038)	(0.042)	(0.048)	(0.053)
Number of previous attempts		0.064	0.062	0.035	0.030
		(0.047)	(0.048)	(0.059)	(0.057)
Work hours per week		−0.002	−0.002*	−0.002	−0.002
		(0.001)	(0.001)	(0.001)	(0.001)
Semester hours completed		0.000	0.000	0.000	0.000
		(0.001)	(0.001)	(0.001)	(0.001)
Current semester hours		−0.000	0.000	−0.004	−0.003
		(0.007)	(0.007)	(0.008)	(0.008)
Number of standardized absences		−0.001	−0.000	−0.004	−0.004
		(0.004)	(0.003)	(0.004)	(0.005)
High-school course in economics		0.036	0.037	0.023	0.023
		(0.022)	(0.024)	(0.023)	(0.026)
Business student		−0.022	−0.020	−0.005	−0.002
		(0.024)	(0.025)	(0.028)	(0.028)
Final exam section mean			0.067*		0.061*
			(0.029)		(0.031)
Final exam section SD			0.279*		0.312**
			(0.130)		(0.125)
Student evaluation section mean			−0.010		−0.022**
			(0.009)		(0.005)
Student evaluation section SD			−0.040		−0.060*
			(0.031)		(0.029)
Observations	263	263	263	235	235
Adjusted R^2	0.06	0.15	0.18	0.30	0.32

Robust standard errors are in parentheses and are adjusted for within-section correlation of errors.
* Significant at 10%; ** significant at 5%; *** significant at 1%.

performance. Time spent at work may reduce time available to study, and we find some evidence that the average number of weekly work hours negatively affects student performance. Current semester hours, previously completed semester hours, and the number of absences are not significant predictors of change in our data, though each has the expected sign.

In Specifications (4) and (6), we include the section-specific means and standard deviations of the common portion of the final exam and student evaluations to serve as instruments for unobserved

instructor differences. As we discuss in section 2, we are unable to directly control for unobserved instructor-level heterogeneity due to the collinearity of our experimental and instructor dummies. We control, in part, for instructor-level effects by including the section-specific controls described above and by allowing for the potential correlation of errors within class sections (which is likely due largely to these unobserved instructor effects). After controlling for student- and instructor-level effects in Specifications (4) and (6), we see that the effect of the experimental approach remains positive and statistically significant in the range of a 2.42–2.59 question improvement differential. Interestingly, coefficients on the distributions of the final exam and student evaluations are often statistically significant. In general, students in sections with higher average final-exam scores also performed better, other things equal, on the change measure. Additionally, students in sections that rated the course and instructor higher on student evaluations faired worse on the change score than students in sections with lower evaluations.

Briefly, results reported in Table 5 using the gap-closing measure support those in Table 4. Specifically, students in the experimental sections gained a significantly higher percentage of the available points on the postcourse TUCE exam, other things equal. The average differential in the gap-closing measure that is attributable to participation in the experimental sections is between 11.1 and 12.3 percentage points, other things equal. Model fit was not as good when using the gap-closing measure; while parameter estimates are often of the anticipated sign, they are generally imprecise.

Results presented thus far demonstrate that students in the experimental sections experienced larger improvements on the TUCE than similar students in the nonexperimental sections. It is unclear at this point, however, in what areas the treatment group experienced these gains. For instance, did experimental students experience larger gains relative to the control group on questions that tested basic concepts, or did these gains come from more difficult questions? Or were there certain content areas, such as international economics or market structures, in which experimental students outperformed their nonexperimental counterparts? Fortunately, TUCE questions are categorized into several content and cognitive categories. Questions can be classified as belonging to one of six different content areas, such as "basic economic problems" or "market failures, externalities, government intervention, and regulation." Similarly, each question can belong to one of three cognitive categories, based on Bloom's taxonomy of educational objectives (Bloom et al. 1956). Moving from lower to higher levels of learning, the TUCE classifies questions as testing either "recognition and understanding" (combining the first two levels of Bloom's taxonomy, knowledge and comprehension), "explicit application" (part of the third level of Bloom's taxonomy, application), or "implicit application" (also part of the third level of Bloom's taxonomy). Given the six content and three cognitive categories, questions can be assigned to one of 18 different content/cognitive domains. Because no questions tested recognition and understanding of international economics, only 17 domains have questions assigned to them.

Table 6 presents estimates of the experimental section dummy variable for each content/cognitive category. Estimates were obtained using the difference in pre- and postcourse TUCE scores as the dependent variable and a specification identical to Specification (4) in Table 4 that employs all of our control variables. Thus, each coefficient reported in Table 6 represents the number of points, on average, experimental students gained on the TUCE relative to the control group, other things equal. The number of questions assigned to each category differs, so Table 6 also reports in brackets the number of questions assigned to each cell. For example, Table 6 reports that the

Table 6. Estimated Coefficients on Experimental Section Dummy across Content and Cognitive TUCE Categories

	Cognitive Categories			
Content Categories	Recognition and Understanding	Explicit Application	Implicit Application	Cumulative Content
The basic economic problem	0.004	0.201**	0.206***	0.410***
	[1]	[1]	[2]	[4]
Markets and the price mechanism	0.108	0.455*	0.131***	0.699**
	[1]	[6]	[1]	[8]
Cost, revenue, profit maximization, and market structure	−0.064	0.094	0.351***	0.394*
	[3]	[2]	[2]	[7]
Market failures, externalities, government intervention, and regulation	0.494***	0.157**	0.271***	0.924***
	[4]	[1]	[2]	[7]
Income distribution and government redistribution	−0.024	−0.005	0.299**	0.267
	[1]	[2]	[3]	[6]
International economics	—	0.141***	−0.031	0.116
		[2]	[1]	[3]
Cumulative cognitive	0.522*	1.082*	1.275***	2.592**
	[10]	[14]	[11]	[33]

Experimental effects are estimated using Specification (4) from Table 4. Numbers in [brackets] are the number of TUCE test items in each content/cognitive category. Note that these numbers do not sum to 33 because two TUCE questions are in multiple categories.

* Significant at 10%; ** significant at 5%; *** significant at 1%.

experimental group gained an average of 0.39 points more than the control group in the "cost, revenue, profit maximization, and market structure" content area, with this difference driven primarily from achievements on questions testing implicit application of these concepts. Note that, when statistically significant, estimated coefficients are positive. Of the 17 assigned content/cognitive areas, only seven contain statistically insignificant coefficients. Though the experimental group achieved significant gains across the various content and cognitive categories, it appears that gains are most evident in implicit application questions (the highest cognitive level) and in the "market failures, externalities, government intervention, and regulation" content area.[20]

As a final step, we grouped the 10 most difficult questions and 10 least difficult questions as defined by the U.S. national norming sample of respondents on the postcourse TUCE.[21] Table 7 reports that students in the experimental group had significant gains relative to students in the control group for both low- and high-difficulty questions. Thus, combined with the results from Table 6, these results suggest that the experimental teaching method yielded benefits across the various content, cognitive, and difficulty levels.

[20] The Bergstrom and Miller (2000) textbook, from which experiments were taken, includes three separate experiments that address market failures, covering pollution externalities, network externalities, and adverse selection.

[21] This information is included in the TUCE examiner's manual. The 10 most difficult postcourse TUCE questions in the national norming sample had successful response rates ranging from 33 to 40%. The 10 least difficult questions had successful response rates ranging from 55 to 77%. Questions in both groups spanned the three cognitive and six content categories.

Table 7. Estimated Coefficients on Experimental Section Dummy across Levels of TUCE Difficulty

	Estimated Differential
Ten easiest questions on postcourse TUCE	0.864**
	[10]
Ten hardest questions on postcourse TUCE	0.793**
	[10]

Experimental effects are estimated using Specification (4) from Table 4. Numbers in [brackets] are the number of TUCE test items in each category of difficulty.
** Significant at 5%.

Differential Effect of Experiments across Student Characteristics

Although we have demonstrated that students exposed to a comprehensive experimental approach have, on average, higher gains in achievement than their nonexperimental cohorts, it is possible that some students in the experimental group would have faired better in the traditional approach. In other words, the efficacy of both teaching methods may be student dependent, a result that would be consistent with learning theory (Fels 1993). Therefore, we estimate the potential differential effect of the experimental approach on a variety of student characteristics. These estimates can provide an indication of which approach may better suit certain types of students. Specifically, Table 8 reports estimates of the differential effect of the experimental treatment across each of our student characteristics.

The main effect of the treatment group is again positive and statistically significant using either of our dependent variables. In contrast with Dickie's (2000) findings on aptitude, Table 8 reports that there is a statistically significant positive relationship between GPA and achievement for the control group, but no significant association between these variables for students in the experimental group.[22] Similarly, using SAT scores instead of GPA, we see that, while higher SAT scores generally have a positive and statistically significant effect on achievement for nonexperimental students, there is no such relationship for students in the experimental sections, with the one exception being that, in the change model, higher verbal SAT scores are shown to have a positive impact on performance for both experimental and nonexperimental students.[23] Additionally, the positive male-gender effect so robustly demonstrated in Tables 4 and 5 and elsewhere in the existing literature is tempered when using the experimental approach. Although males have a significant achievement advantage over females in the traditional sections, this gender advantage disappears for students in the experimental sections. Finally, nonwhites in experimental sections appear to be at a significant disadvantage to their nonexperimental counterparts.[24]

Thus, our results indicate that certain groups of students that have historically been at a disadvantage in economics classes, particularly low-ability/achievement students and females, appear to perform much better relative to appropriate comparison groups when they participate in the

[22] In other words, summing the estimated coefficients of the main effect of GPA and the interaction of GPA with the experimental dummy variable yields a number that is statistically indistinguishable from zero.

[23] In other words, although we do not estimate a significant differential effect for the experimental group for any of the SAT scores, the sum of the main effect of SAT and the interaction of SAT with the experimental dummy yields a statistically insignificant effect for all cases except verbal SAT scores using the change model.

[24] Again, although we do not, in general, estimate statistically significant main effects or interactions for the nonwhite variable, the linear combination of the main effect for nonwhite and its interaction with the experimental dummy variable is negative and statistically significant, indicating that nonwhites in the experimental group do worse than similar students in the control group.

Table 8. Important Interactions with Student Characteristics

	Difference in TUCE		Gap-Closing Measure	
Independent Variables	(1)	(2)	(1)	(2)
Intercept	−14.939	−29.279	−0.740	−1.240
	(21.800)	(26.209)	(0.985)	(1.165)
Experimental section	19.169***	22.131***	0.813***	0.873**
	(2.456)	(5.293)	(0.130)	(0.284)
Precourse TUCE	−0.481***	−0.685***		
	(0.042)	(0.072)		
GPA	1.792**		0.068**	
	(0.600)		(0.027)	
GPA × experimental section	−2.054*		−0.090*	
	(1.090)		(0.047)	
SAT math		0.011**		0.000*
		(0.004)		(0.000)
SAT math × experimental section		−0.010		−0.000
		(0.007)		(0.000)
SAT verbal		0.021***		0.001
		(0.004)		(0.000)
SAT verbal × experimental section		−0.005		−0.000
		(0.010)		(0.000)
Male	1.894***	1.451***	0.074***	0.053***
	(0.434)	(0.416)	(0.020)	(0.014)
Male × experimental section	−1.815**	−2.311*	−0.058**	−0.059
	(0.618)	(1.031)	(0.025)	(0.041)
Nonwhite	−0.778	−0.808	−0.040	−0.039
	(0.885)	(1.174)	(0.050)	(0.064)
Nonwhite × experimental section	−1.848*	−1.792	−0.082	−0.082
	(0.882)	(1.430)	(0.051)	(0.076)
Number of previous attempts	1.545	0.815	0.027	−0.027
	(1.262)	(1.273)	(0.048)	(0.049)
Number of previous attempts	4.178**	3.403	0.203***	0.211**
× experimental section	(1.495)	(2.013)	(0.054)	(0.072)
Work hours per week	−0.038	−0.036	−0.001	−0.001
	(0.030)	(0.032)	(0.001)	(0.001)
Work hours per week	−0.162**	−0.112	−0.006*	−0.004
× experimental section	(0.068)	(0.085)	(0.003)	(0.004)
Semester hours completed	0.012	0.020	0.001	0.001
	(0.033)	(0.025)	(0.002)	(0.001)
Semester hours completed	−0.065	−0.073*	−0.002	−0.002
× experimental section	(0.044)	(0.039)	(0.002)	(0.002)
Current semester hours	0.080	0.081	0.002	0.002
	(0.149)	(0.153)	(0.008)	(0.008)
Current semester hours	−0.567**	−0.693**	−0.025**	−0.031*
× experimental section	(0.203)	(0.280)	(0.010)	(0.014)
Number of standardized absences	0.015	−0.126	0.001	−0.004
	(0.089)	(0.124)	(0.004)	(0.005)
Number of standardized absences	−0.375**	−0.099	−0.017**	−0.007
× experimental section	(0.137)	(0.152)	(0.006)	(0.006)
High-school course in economics	0.917	0.757	0.026	0.023
	(0.573)	(0.752)	(0.025)	(0.032)

Table 8. Continued

Independent Variables	Difference in TUCE		Gap-Closing Measure	
	(1)	(2)	(1)	(2)
High-school course in economics	1.424	1.721	0.073	0.060
× experimental section	(1.063)	(1.330)	(0.044)	(0.055)
Business student	−0.504	0.011	−0.027	−0.010
	(0.646)	(0.843)	(0.029)	(0.035)
Business student ×	1.983*	2.150	0.073	0.075
experimental section	(1.018)	(1.220)	(0.052)	(0.057)
Final-exam section mean	0.810*	1.278*	0.035*	0.051*
	(0.404)	(0.601)	(0.017)	(0.023)
Final-exam section SD	2.009**	4.843*	0.121	0.231*
	(0.068)	(2.39)	(0.094)	(0.110)
Student evaluation section mean	−0.050	−0.274**	−0.003*	−0.014*
	(0.061)	(0.099)	(0.001)	(0.006)
Student evaluation section SD	−0.019	−0.599*	−0.013	−0.039*
	(0.019)	(0.293)	(0.007)	(0.020)
Observations	263	235	263	235
Adjusted R^2	0.25	0.38	0.20	0.35

Robust standard errors are in parentheses and are adjusted for within-section correlation of errors.
* Significant at 10%; ** significant at 5%; ***significant at 1%.

experimental curriculum. This "leveling of the playing field" does not appear to hold, however, for nonwhites in the experimental group, who appear to perform worse than their nonexperimental peers.

In addition to the differentials estimated for GPA, SAT, gender, and ethnicity, results also indicate that students who worked more (at employment) or who carried more class hours, students who had completed more total semester hours, and students who were absent from class more often all performed relatively worse than similar students as a result of the experimental treatment. On the other hand, we estimate positive differentials resulting from the experimental pedagogy for students who had more previous attempts at the course, who had high-school economics, and who were business majors.

It is argued that the experimental approach provides a framework (i.e., learning schema) within which students place the theories presented in a principles course. Such a structure would likely benefit students who had previous exposure to the basic concepts (e.g., students who had previously attempted the course or taken high-school economics) or who tend to benefit from experience and examples (e.g., females and business majors). However, the development of the framework and linking of theoretical concepts using experimental techniques is time intensive. As a result, it is not surprising to us that students with less time to devote to the class (i.e., students with heavier employment or course loads or greater absences) would perform better in the traditional lecture-oriented framework than under the experimental treatment.

4. Conclusion

In our study of 300 microeconomics principles students, we find that those students who were taught using an experimental approach experienced a significantly larger improvement on the Test of Understanding in College Economics (TUCE) than similar students taught with the traditional lecture-

oriented approach. After controlling for student- and section-level characteristics, our results indicate that students in the experimental sections improved their TUCE score by an average of 2.42–2.99 questions over the control group. Our results are robust to potential issues of positive selection bias, endogeneity of precourse ability, and censoring of the dependent variable. In addition to student learning measured by changes in pre- and postcourse TUCE scores, we also report other outcomes, including performance on a departmental final exam, student evaluations, and class attrition rates. We find, however, little difference between the two groups for these other outcomes. Finally, we find that certain student characteristics, like GPA, SAT scores, gender, and major, can affect the likelihood of achievement in an experimental course.

While our results are robust, we acknowledge the relative specificity of our data. Additional research on the effect of the experimental pedagogy on student learning should be undertaken. We recommend multiuniversity studies, and we suggest that the experimental approach be studied using a variety of class sizes. We further call for the use of research designs that allow for a more explicit accounting of unobserved instructor effects, realizing of course the considerable resources that such designs might require.

Longer term outcomes are also of interest. These include effects on students' choices of major, enrollment and performance in additional economics courses, and long-term differentials in retention of the material. The longitudinal data required for such studies requires following a cohort of students through graduation and beyond, and like our other recommendations, is resource intensive. We plan to follow our cohort of 300 students through graduation to collect data on these additional outcomes, which we will report in future work.

References

Becker, William E. 1982. The educational process and student achievement given uncertainty in measurement. *American Economic Review* 72:229–36.

Becker, William E. 1997. Teaching economics to undergraduates. *Journal of Economic Literature* 35:1347–73.

Becker, William E. 2001. What does the quantitative research literature really show about teaching methods? Unpublished paper, Indiana University.

Becker, William, Robert Highsmith, Peter Kennedy, and William Walstad. 1991. An agenda for research on economic education in colleges and universities. *American Economic Review Papers and Proceedings* 81:26–31.

Becker, William E., and John R. Powers. 2001. Student performance, attrition, and class size given missing student data. *Economics of Education Review* 20:377–88.

Becker, William E., and Michael K. Salemi. 1977. The learning and cost effectiveness of AVT supplemented instruction: Specification of learning models. *Journal of Economic Education* 8:77–92.

Bergstrom, Theodore C., and John H. Miller. 2000. *Experiments with economic principles: Microeconomics*. 2nd edition. Burr Ridge, IL: McGraw-Hill Higher Education.

Bloom, B. S., M. D. Engelhart, E. J. Frost, W. H. Hill, and D. R. Krathwohl. 1956. *Taxonomy of educational objectives. Handbook I: Cognitive domain*. New York: David McKay.

Cardell, N. Scott, Rodney Fort, Wayne Joerding, Fred Inaba, David Lamoreaux, Robert Rosenman, Ernst Stromsdorfer, and Robin Bartlett. 1996. Laboratory-based experimental and demonstration initiatives in teaching undergraduate economics. *American Economic Review Papers and Proceedings* 86:454–9.

Chizmar, John F., and Anthony L. Ostrosky. 1998. The one-minute paper: Some empirical findings. *Journal of Economic Education* 29:3–10.

DeYoung, Robert. 1993. Market experiments: The laboratory versus the classroom. *Journal of Economic Education* 24:335–51.

Dickie, Mark. 2000. Experimenting on classroom experiments: Do they increase learning in introductory microeconomics? Unpublished paper, University of Southern Mississippi.

Durden, Garey C., and Larry V. Ellis. 1995. The effects of attendance on student learning in principles of economics. *American Economic Review Papers and Proceedings* 85:343–6.

Fels, Rendigs. 1993. This is what I do, and I like it. *Journal of Economic Education* 24:365–70.

Fleisher, Belton, Masanori Hashimoto, and Bruce Weinberg. 2002. Foreign GTAs can be effective teachers of economics. *Journal of Economic Education* 33:299–325.

Frank, Bjorn. 1997. The impact of classroom experiments on the learning of economics: An empirical investigation. *Economic Inquiry* 35:763–9.

Gramlich, Edward M., and Glen A. Greenlee. 1993. Measuring teaching performance. *Journal of Economic Education* 24:3–13.

Gremmen, Hans, and Jan Potters. 1997. Assessing the efficacy of gaming in economic education. *Journal of Economic Education* 28:291–303.

Kennedy, Peter E., and John J. Siegfried. 1997. Class size and achievement in introductory economics: Evidence from the TUCE III data. *Economics of Education Review* 16:385–94.

Mankiw, N. Gregory. 2001. *Principles of economics*. 2nd edition. Stamford, CT: Dryden Press.

Maxwell, Nan L., and Jane S. Lopus. 1994. The Lake Wobegon effect in student self-reported data. *American Economic Review Papers and Proceedings* 84:201–5.

Raimondo, Henry J., Louis Esposito, and Irving Gershenberg. 1990. Introductory class size and student performance in intermediate theory courses. *Journal of Economic Education* 21:369–81.

Shmanske, Stephen. 1988. On the measurement of teacher effectiveness. *Journal of Economic Education* 19:307–14.

Siegfried, John J., and Rendigs Fels. 1979. Research on teaching college economics: A survey. *Journal of Economic Literature* 17:923–69.

Watts, Michael, and William Bosshardt. 1991. How instructors make a difference: Panel data estimates from principles of economics courses. *Review of Economics and Statistics* 73:336–40.

Williams, Arlington, and James Walker. 1993. Computerized laboratory exercises for microeconomics education: Three applications motivated by the methodology of experimental economics. *Journal of Economic Education* 24:291–315.