Research in Economic Education

In this section, the *Journal of Economic Education* publishes original theoretical and empirical studies of economic education, dealing with the analysis and evaluation of teaching methods, learning, attitudes and interests, materials, or processes.

PETER KENNEDY, Section Editor

Additional Evidence on the Relationship between Class Size and Student Performance

J. J. Arias and Douglas M. Walker

Abstract: Much of the economic education literature suggests that the principles of economics class size does not significantly affect student performance. However, study methods have varied in terms of the aggregation level (student or class), the measure of performance (TUCE or course letter grade), and the class size measure (e.g., students who completed both the TUCE pretest and posttest). The authors perform an experiment with principles students using total exam points as the dependent variable in a model to explain student performance. By using the same instructor for all sections, the authors control variation in instruction, lecture material, and topic coverage; they also account for variation in student abilities. In contrast to many other studies, the authors find statistically significant evidence that small class size has a positive impact on student performance.

Key words: class size, education, testing
JEL codes: A22, I21

Students, parents, and administrators often believe that smaller university classes are better.[1] It is probably true that students in smaller classes receive more personal attention, are more likely to be known by name by their professors, and may feel more comfortable to ask questions or otherwise participate in class

J. J. Arias *is an assistant professor of economics, and* ***Douglas M. Walker*** *(e-mail: doug.walker@gcsu.edu) is an associate professor of economics, both are at Georgia College & State University.* *The authors are grateful to Esenç Balam, John Jackson, Peter Kennedy, David Laband, John Swinton, and three anonymous referees for helpful comments.*

discussions. But do students really learn more in small classes? Much of the evidence in the economic education literature suggests the answer to this question is, "No."[2]

Various methods have been employed in studying the effects of class size on the performance of economics students. Because each study comes with its own caveats, there is still a need for more evidence and debate. Our purpose in this article is to offer some additional evidence on how class size and other factors affect learning in the survey of economics class. We conducted a controlled experiment in which the same instructor taught all the sample sections, delivered the same lectures, and used the same exams and grading mechanisms. The only major difference among the classes was their size.

LITERATURE REVIEW

The intuition behind the idea that smaller classes are better learning environments is well known. In smaller economics classes, interactive discussions may be used more than lectures, facilitating better "delayed recall" learning (McKeachie 1990, 190–91) and critical thinking (Raimondo, Esposito, and Gershenberg 1990, 371–72). In smaller classes, instructors know the students' names, and students may not want to disappoint an instructor who knows them personally (Siegfried and Kennedy 1995, 347, note 1). Lippman (1990, 193) boldly claimed that the class size debate is settled: "There is a strong relationship between class size and student achievement. [This is] in accord with common sense, intuition, and anecdotal observation. Small classes are better."[3]

Siegfried and Kennedy (1995) examined data from the Test of Understanding College Economics (TUCE III) and discussed the issue of pedagogy and class size. They suggested that there are not significant differences in how different class sizes are taught, at least for introductory economics classes. This finding is not surprising when one considers that many faculty members teach principles classes each semester. It seems unlikely that they would adjust the class format each semester simply because of differing class sizes. Siegfried and Kennedy (349–50) confirmed this with evidence from their sample. The fact that "instructors do not adjust their teaching methods to class size" (p. 347) suggests that class size should not affect learning.[4]

Kennedy and Siegfried (1997) presented the most comprehensive analysis of class size in economics. They used the TUCE data for almost 200 classes at 53 different universities to test a GLS (generalized least squares) model explaining scores on the TUCE posttest.[5] As in their previous work, they concluded that "class size does not affect student achievement" and further that "class characteristics over which instructors or department chairs have control also do not influence achievement" (p. 385).

Siegfried and Walstad (1998) provided a recent and comprehensive review of the literature on teaching economics. In discussing the evidence on class size, they concluded that size does not seem to matter, once it rises above 20 students. Still, they indicated the need for more research, for example, comparing performance at a variety of different class sizes.[6]

Contrary to this evidence, Lopus and Maxwell (1995) used the TUCE data but reported a significant *positive* relationship between achievement and class size. They explained, "This could indicate that large classes are efficient, that higher quality institutions in the TUCE III sample offer larger classes, [that] higher quality instructors are assigned to teach larger classes" (p. 348), or that students in larger classes are more experienced in taking multiple-choice exams.[7] Kennedy and Siegfried (1997, 387) were critical of the Lopus and Maxwell result. Obviously, there is disagreement about the proper method, even among studies using the TUCE data.

Most recently, Becker and Powers (2001) provided important evidence that the relationship between class size and student performance is sensitive to *which* measure of class size is chosen. Referring to studies using the TUCE data, they explained:

> Contrary to studies that have used an average or an end-of-term class size measure and find no class-size effect, beginning class size is found to be significant and negatively related to learning, all else equal. In part, this is the result of students in larger classes being significantly more likely than students in smaller classes to withdraw from the course before taking the posttest. (p. 378)

This result helps to explain why so many researchers who use TUCE data have found no significant relationship between class size and student performance: Many of the poorly performing students withdraw from the class and are not included in the test samples. We discuss this important issue, and how it relates to our study, later in this article.

Several other authors offer useful analyses related to principles class performance. Laband and Piette (1995) examined students who took the principles courses at a community or junior college and then transferred to a university. They found that the transfer students tended to perform worse than the nontransfer students in the upper-level economics coursework. After controlling for a number of variables, they concluded that the instruction is probably worse at the community colleges.

Raimondo, Esposito, and Gershenberg (1990) used intermediate course grade as a dependent variable in examining the effect of introductory class sizes on students' performance in intermediate-level courses. They found that smaller principles classes have a positive effect on intermediate macro performance. No significant relationship appeared to exist within the micro sequence of courses.

Hancock's (1996) experiment was similar to ours. He tested three large and six small sections of statistics classes, holding constant instructor, lectures, and exams. Despite this similarity to our experiment, Hancock's empirical analysis was quite limited, as it did not consider factors other than class size that may affect student achievement. In any case, Hancock found no significant difference in the performance among the students in different size sections. He noted,

> If learning experience is not demonstrably harmed by significant increases in enrollment caps, then it is certainly harmed by not increasing them. Resources lost to staffing unnecessary sections are opportunities denied in instructional development

and technology—upgrades in skills, support, and facilities. A one-size-fits-all policy may also be a disservice to courses where a smaller class size could genuinely make a difference. Treating all courses the same may be one of the most expensive and most common mistakes in higher education. (p. 481)

The insignificance of class size in statistics course performance does not, as Hancock admited, necessarily imply class size is insignificant in other disciplines, or other classes within a discipline. Maybe class discussion, for example, is more effective and useful in economics than it is in statistics.

THE EXPERIMENT AND DATA STATISTICS

Perhaps the most salient feature of the literature on class size is the variety of methods employed by researchers. To measure performance or learning, some researchers have relied on standardized test scores, whereas others use course letter grades. The sample sizes of studies in the literature vary greatly. Also, there is inconsistency in the types of explanatory variables used in the various studies. Hence, one must be careful comparing these studies and generalizing from their results.

The goal in designing our experiment was to isolate the effect of class size on student performance by holding constant as many variables as possible. We purposely made the class sizes different, of course, but all other aspects of the classes were controlled. We believe this experiment to be unique, at least in the literature with which we are familiar.[8]

The Experiment Class Sections

Our university is a public liberal arts college with nearly 5,500 students. Normally, classes at the university are no larger than 50 students. Although the students have opted to attend a liberal arts college, the students here are similar to students at large state universities at which we have taught.[9] During the 2000–2001 academic year, we were given administrative permission to run an experiment on class size using Georgia College & State University's (GC&SU) course, economics and society. This is the principles survey course required for all nonbusiness majors and is typically taken during the freshman or sophomore year.[10]

We organized four experimental sections of economics and society during the year, two sections per semester. Each semester we had one large section (maximum of 89 students) and one small section (maximum of 25 students). We scheduled the sections back-to-back at popular mid-morning times on the same days. This was intended to minimize the effect class time might have on students' class selection.

Enrollment Procedure

At our university, students are able to enroll, online or in person, well in advance of the beginning of the semester. For fall classes beginning in August, students begin enrolling in March; for the spring classes beginning in January, students begin enrolling in October. In deciding which courses to take, students

use an online system that shows them information about the classes offered, including days and times, professor's name, classroom, maximum class size, and number of students enrolled in each section.

There is priority registration, but only for seniors (i.e., students with at least 90 semester hours of credit). The first 24 hours of the registration period are reserved for these students so that they have a better chance of getting courses they need to graduate. This reserved registration time is irrelevant to our experiment, however, because the overwhelming majority of the students in economics and society are freshmen and sophomores.

The registration process does raise a potential sample selection bias problem, which we address later.

Measuring Achievement/Learning

Similar to Hancock (1996), Raimondo, Esposito, and Gershenberg (1990) and Watts and Lynch (1989), we used the students' course performance, rather than the TUCE, as the measure of output.[11] However, our study was different because we used total exam points earned,[12] rather than course letter grade.[13] The students took four multiple-choice exams with 30 questions each, for a total of 120 possible points.[14]

Like Hancock (1996), we tried to minimize the amount of variation among the lectures. The different class sections in our experiment were all taught by the same instructor (Arias), who used the same textbook, assignments, lectures, exams, and grading procedures in all sections. The only intentional difference among the classes was their size. We collected data on students' performance and a variety of other variables, using official university sources (discussed later).

Exam Security

Although the same exams were used in all the experiment sections, students were not allowed to keep the exams nor were the exams handed back to the students after grading. Hence, we were fairly certain that students did not get copies of the exams prior to taking them in class. It is possible that students in the earlier section could have shared information about the exams with the students in the later section. However, we seriously doubt this occurred, for three reasons. First, the courses were scheduled back-to-back, which left little time for such interaction, because the exams in the two sections began only one hour apart. (The fall semester sections were at 9 and 10 a.m. In the spring semesters, they were at 10 and 11 a.m. During both semesters, the large class was scheduled first.) Second, the course sections were taught in different buildings, which would make it difficult for students to share information about the exams.[15] Still, one may suspect that the large-section students could have given information to the small-section students. To test this, we examined the final exam scores alone. Although the large classes wrote their interim exams one hour before the small sections (consistent with the class times), the small sections wrote their final exams *two days before* the large sections. We re-ran the regression,

using final exam score rather than total exam points as the dependent variable. Even though the small sections wrote their final exams first, their scores were still significantly higher than those of the students in the large sections. This result suggests that cheating was not the source of our positive small-class-size effect.

Statistical Analysis

Prior to developing an econometric model to explain the effect of class size on the students' performance, we present some descriptive statistics of our sample and examine the relationship between small and large class sizes.

The class size data were derived by taking the number of students who were assigned grades (A, B, C, D, F, or W) for the course and making two adjustments. First, students who withdrew (received W grades) from the course were eliminated from the sample. There were 18 students who withdrew from these sections, 15 from large sections, and 3 from small sections. According to Becker and Powers' (2001) evidence, these students would likely have performed worse than students who remained in the course. Our evidence supports this, as these students' average grade, prior to their withdrawal, was about 12 percent lower than the remaining students' final mean score. The second adjustment was to drop from the sample the 8 students who did not take the final exam. All of these students were in the large sections and failed the course. These two adjustments left us with a total of 195 observations for the statistical analysis, as indicated in Table 1, where 0 and 1 subscripts refer to the large and small sections, respectively, and *f* and *s* represent fall and spring semesters.

Becker and Powers (2001) found that dropping students who withdrew will bias the results. In our case, dropping the students who withdrew or did not take the final exam resulted in an overstatement of the large class performance. This was because the large sections had disproportionately more of these students. We will discuss this bias and its effect on our results later.

TABLE 1. Class Size Data

	Fall large	Fall small	Spring large	Spring small	Total
Students assigned grades	86	26	84	25	221
Withdrawals	–7	–3	–8	–0	–18
No final exam	–3	–0	–5	–0	–8
Net students included in statistical analysis	$n_{f0} = 76$	$n_{f1} = 23$	$n_{s0} = 71$	$n_{s1} = 25$	195
No SAT/ACT scores	–7	–0	–7	–1	–15
Net students included in regression analysis	69	23	64	24	180

Note: Subscripts 0 and 1 refer to the large and small class sizes, respectively.

TABLE 2. Total Point and Percentage Means and Variances by Class Section

Class	Sample mean points $\bar{X}$ and variance (S^2)	Mean score, percentage of 120 pts.
Large, fall ($_{f0}$)	76.1 (196.36)	63.4
Large, spring ($_{s0}$)	80.8 (199.14)	67.3
Large, pooled	78.4 (201.86)	65.3
Small, fall (*f1*)	84.3 (134.13)	70.4
Small, spring (*s1*)	88.7 (161.46)	73.9
Small, pooled	86.6 (150.20)	72.2

Statistics on the classes' point total means and variances, along with mean percentages, are shown in Table 2. Tests showed that the sample variances and means were not significantly different between the small classes,[16] and the variances of the large classes were equal; however, the means of the large sections were weakly unequal.[17] Nevertheless, we assumed that the data for each of the four sections were drawn from a normal distribution and were from the same population. The obvious advantage to pooling the samples was that it yielded more degrees of freedom and more efficient parameter estimates. The pooled point total means and variances are also indicated in Table 2.

We now come to the main question of interest: Are the mean scores equal across class size? We had to first decide whether or not to assume a common variance. The hypothesis could be expressed as $H_0: \sigma_1^2/\sigma_0^2 = 1$. The F statistic was 1.34 (p value = 0.121). We failed to reject the null hypothesis of equal variances across class sizes. The pooled estimate for the common variance was $S^2 = 189.28$.

Finally, we tested for equality of means between the pooled large and pooled small class scores ($H_0 : \mu_1 - \mu_0 = 0$). The t test statistic was 3.60 (p value = 0.0004), indicating the means across class size were statistically different. From Table 2, it appears the mean was greater in the small classes. Indeed, we could not reject the hypothesis of a larger mean in the small classes: for $H_0 : \mu_1 - \mu_0 > 0$, the t statistic = 3.60 (critical value with $\alpha = 0.01$ is –2.36; p value = 0.999).[18]

This statistical evidence that the smaller class score mean is higher suggests smaller classes are more conducive to learning economics. However, this analysis is incomplete because it does not explain the source(s) of the difference in means. In the next section, we posit an econometric model to explain student performance.

ECONOMETRIC MODEL AND RESULTS

We collected data on several variables that we expected could help explain student performance in the economics survey course and tested this OLS (ordinary least squares) model:

$$SCORE = \beta_1 + \beta_2\, FEMALE + \beta_3\, FRESHFALL + \beta_4\, GPA + \beta_5\, SATM + \beta_6\, SATV + \beta_7\, SMALL + \beta_8\, TRANSFER + \beta_9\, YEARBORN + \varepsilon.$$

The dependent variable, *SCORE*, is the students' point totals for the four exams. This score can range from 0 to 120.

Several measures of the students' academic abilities were included as explanatory variables. The student's SAT math and verbal scores (*SATM* and *SATV*, respectively) are typically interpreted as signals (or predictors) of the student's academic potential. For the few students who took the ACT instead of the SAT, we converted their scores to the SAT equivalents.[19] We did not have SAT or ACT data on 15 students, so they were omitted from the study, leaving 180 observations for the regression analysis (Table 1).[20]

The GPA (grade point average) represents evidence of the students' performance. The *GPA* used in the model is the student's cumulative college GPA, according to official university data, not student-reported GPAs. The *GPA* statistic was calculated after the term in which the student took the economics course. For the fall semester economics students, the *GPA* measure was taken in December 2000, after final grades for the fall semester were reported. For spring students, the *GPA* measure was taken in May 2001, after spring grades were reported. We removed the effect of the economics course grade from the *GPA* calculation, so that it was not affected by the economics course grade or the *SCORE* variable in the model.

A variety of other variables was included in the model. Most of the students taking economics and society were either freshmen or sophomores. All students must complete a common 2-year core, and as a result, we expected most students taking this course to be at similar stages in their academic careers. Nevertheless, we included the students' year of birth (*YEARBORN*) to account for older students possibly being more mature (see Laband and Piette 1995, 337). Finegan and Siegfried (1998) used a dummy for freshmen students. We included *FRESHFALL* as a dummy for freshman students taking the course during the fall semester. We expected first-semester freshmen to perform worse in this course because economics may be more demanding than their high school courses or the other courses they take during their first semester at college.[21]

Class size is represented by a dummy variable, *SMALL* (1 if the student enrolled in a small section; 0 for the large classes). We used two other dummies in the model; *FEMALE* represents student sex (1 if female; 0 if male). Following Laband and Piette (1995), we tried to account for the effect of transferring on performance (*TRANSFER*). This takes a value of 1 if the student transferred any credit hours to GC&SU, 0 if not. The error term is ε, which we assumed was normally distributed. The regression results are displayed in Table 3.

As Table 3 indicates, we found that small class size does have a positive impact on class performance (significant at the .05 Type I error level). Students in the small classes (average size of 23.5 students) scored almost 4 points or 3 percent higher than students in the large classes (average size of 66.5 students). Unsurprisingly, students with higher GPAs and SAT math scores performed better in the economics class.

Freshmen taking the course in the fall term appeared to perform significantly worse, on average by about 4.7 points (3.9 percent), than nonfreshmen and freshmen

TABLE 3. Regression Results (Dependent Variable = *SCORE*)

Variable	Coefficient	*t* statistic
Constant	26.703	0.536
FEMALE	–1.198	–0.687
FRESHFALL	–4.717**	–2.888
GPA	12.152**	9.104
SATM	0.036**	2.689
SATV	0.004	0.354
SMALL	3.828*	2.125
TRANSFER	0.119	0.069
YEARBORN	0.011	0.018

*Significant at the .05 Type I error level; **significant at the .01 Type I error level.

taking the course during the spring.[22] Students may face some difficulty adjusting to college life or the economics class, and so they perform worse if they take the class their first semester.

Other variables appeared not to affect student performance, including sex, SAT verbal score, transferring, and age.[23] Overall, the model explained about 50 percent of the variation in students' grades ($\bar{R}^2 = 0.49$).

DISCUSSION

In light of the fact that our study varied from most others in terms of design and empirical results, several issues warrant discussion. We first discuss the potential limitations of our study, including the possibility of sample selection bias in our experiment. Next, we attempt to justify our experiment and analysis relative to studies using the TUCE score as a measure of learning and studies that use course letter grade as the dependent variable measuring learning. Our intent was not to criticize previous studies, but rather, to show that more evidence on class size is useful because of the varied methods that have been used in class size analyses.

Sample Selection Bias

During most semesters, students have about eight sections of economics and society to choose from. Our experiment dealt with two sections each semester, and we expected a random distribution of students in the various classes. However, there was a potential sample selection problem related to how students enroll in the courses. There were two specific concerns. First, suppose better students have a preference for smaller class sizes. Second, even if students did not care about class size, suppose the more "eager" or better students tended to enroll sooner than other students. In either of these cases, a sample selection bias would exist, whereby the smaller sections in our experiment would be filled

disproportionately by eager or better students. Perhaps *this* is the cause of our positive small class size effect.[24]

Unfortunately, this issue has not been addressed in previous class-size studies with one exception. Raimondo, Esposito, and Gershenberg (1990) acknowledged that selection bias may have occurred in their experiment, but they dismissed the problem.

> Students were free to select either the large or the small lecture format of introductory economics. This process may have led to a selection-bias problem in drawing the sample of students for this study. However, the selection bias, if it did exist, may be minor. Having worked on registration, the authors were always surprised at the number of students who were either unaware of the different formats in introductory economics or who did not seem to care about the difference. Although there were students who selected a section based on format, other factors, such as time of the class, work schedules, conflicts with other courses, and arrival on and departure from campus, seemed as important (if not more important) in the registration process as class format. (p. 370)

We agree with their argument and attempt to provide evidence to support it.

Of course, the best way to address this issue would have been to prevent it. This could have been done at the initial stage of the experiment when the class schedule was formulated, prior to student registration. Our economics survey class sections are typically 50 students each. On the first day of class we could have combined two of these sections to form a 100-student (large) section. For the small section, we could have split a 50-student section into two smaller sections and used one of them as the experiment's "small" section. As long as the small section students were split randomly, then the sample selection bias issue would have been avoided. (Future studies should be careful to deal with this issue.)

Because we cannot turn-back time, we must use corrective rather than preventative measures. First, to address the possibility that better students tend to prefer smaller class sections, we gave a student survey asking the relative importance of "class size" in choosing a class section. Second, to address the possibility that more eager students enroll sooner than other students, we re-ran the model to include an inverse Mills ratio. After accounting for these concerns, the small class size effect remained positive and significant.

Student surveys. One way of determining whether or not students have a preference for small classes is simply to ask them. During the summer 2003 term, we distributed a short survey to many of the economics principles classes at GC&SU.[25] Although these students are obviously not the same ones in our experiment, they were all in relatively small classes (between 10 and 26 students). We have no reason to believe that these students were much different from the students in our experiment during 2000–01.

The anonymous survey asked students to rate the importance of factors which they may consider in deciding in which section of a particular required course to enroll. The factors listed on the survey included available seats, class size, class time/days, fits in schedule well, professor, recommendation from other students, and a write-in option. The survey included an example from our university's

registration system, which lists the course information. The students were to rate each of the factors from 1 (*not at all important*) to 5 (*extremely important*). Finally, we asked for the students' GPAs.

Of the 72 students surveyed, 54 (or 75 percent) gave "class size" their lowest rating (or tied for lowest rating). Only 3 students (4 percent) gave class size their highest ranking.[26] There appeared to be no systematic relationship between GPA and the perceived relative importance of class size.[27] We saw no evidence from the survey to suggest that the better students sought small course sections. This is consistent with Raimondo, Esposito, and Gershenberg (1990, 370); we found other factors to be more important to students in choosing a class section.

Inverse Mills ratio.[28] The other selection bias concern applies if students are choosing course sections randomly. If the more eager students enroll sooner, then the smaller course sections will fill up sooner simply because their capacities are lower. The result will be a higher proportion of the eager students in the small sections. If these are the better students, then perhaps the positive class-size effect is caused by this sample selection bias. Becker and Powers (2001, 381) noted that "past research, as well as intuition suggest that aptitude and motivation to learn, as measured by grade point average, should have a positive effect [on learning] unless students are selectively or incorrectly reporting their grade point average." Hence, we expected that eagerness would be captured in the students' GPAs. To the extent this was the case, our initial equation already accounted for this "eagerness" effect.

In examining previous analyses that relied on the TUCE data, most of which found no class-size effect, Becker and Powers (2001) found that many of the studies used flawed class-size measures. Specifically, the studies often omitted from the analysis students who either did not complete the TUCE posttest or the questionnaire. Omitting these students introduces a bias in favor of large-class performance because these students tend to be poorer performers than the students who take the posttest and questionnaire, and a disproportionate share of the withdrawn students are in the larger sections.

Using maximum likelihood estimation (MLE), Becker and Powers (2001) introduced an inverse Mills ratio to correct for the sample selection bias associated with attrition. (Also see Becker and Walstad 1990.) After making the correction, they found a positive small class size effect on student performance. We used a similar method to test for the possibility of a sample selection bias associated with how students enroll in the economics courses.

We used the Heckman (1979) two-step method to introduce an inverse Mills ratio (λ), or hazard rate, to model the possibility that eager students disproportionately populate the small class sections. In the first step, a probit model was used to estimate the probability of a student enrolling in a small section. The predicted z values ($\hat{z}$) from the probit were used to estimate the values for the inverse Mills ratio, $\lambda = f(\hat{z})/1 - F(\hat{z})$, where $f(\hat{z})$ is the standardized normal density function evaluated at $\hat{z}$, and $F(\hat{z})$ is the cumulative density function evaluated at $\hat{z}$. In the second step, OLS was used to estimate the coefficients for the original explanatory variables and λ (Greene 2003, 784–87; Kennedy 2003, 297; or Maddala 1983, 231–34). Significance of the coefficient on λ would indicate the presence of selection bias.[29]

The probit model we used to generate λ uses *SMALL* as the dependent variable, with the explanatory variables from the original OLS model[30] (Table 4). If our small class size effect was primarily a result of better students selecting the small sections, or of eager students registering sooner, then we would expect the significance of the *SMALL* variable to disappear from the adjusted OLS model.

In Table 5, we present the results of our original OLS model (Table 3, col. 1) and of incorporating the inverse Mills ratio (λ) (col. 2). The coefficient on λ is insignificant at any conventional level,[31] implying that no statistically significant biases in the original estimates arose from sample selection bias. However, some differences in qualitative inferences result from adding λ. In particular, the *SATM* and *FRESHFALL* variables are no longer statistically significant. More important, though, these differences do not extend to the effects of GPA and class size on grades. Adding the inverse Mills ratio did not alter the positive and significant relationship between *GPA* and *SCORE*, nor did it alter the positive and significant relationship between *SMALL* and *SCORE*. This last result is a key finding of this study.

Generalizing Beyond our Sample

A significant limitation of this study is that the experiment was confined to our university and to nonbusiness economics survey classes. It would be useful for other researchers to perform similar experiments at a variety of universities (with different sizes, selectivity, etc.) and on business economics principles classes. In the conclusion, we offer some suggestions for future research in this area.

Despite the limitations of our study, we believe our methodology has advantages over studies that use the TUCE data or course letter grade as the dependent variables. These issues are discussed below.

Potential Advantages of Using Individual Student Data rather than Class Data

The strongest evidence on the effects of class size on student performance comes from studies using the TUCE data, such as Kennedy and Siegfried (1997), Lopus and Maxwell (1995), and Siegfried and Kennedy (1995). These studies

TABLE 4. Probit Results (Dependent Variable = *SMALL*)

Variable	Coefficient	*z* statistic
Constant	3.525	0.512
FEMALE	0.175	0.732
FRESHFALL	–0.220	–0.975
GPA	0.348*	1.750
SATM	0.005**	2.757
SATV	–0.003*	–1.641
TRANSFER	–0.327	–1.351
YEARBORN	–0.078	–0.899

*Significant at the .10 Type I error level; **significant at the .01 Type I error level.

TABLE 5. Original and Adjusted Regression Results (Dependent Variable = *SCORE*)

Variable	Original model coefficient (*t* statistic)	Adjusted model coefficient (*t* statistic)
Constant	26.703	0.383
	(0.536)	(0.006)
FEMALE	–1.198	–2.20
	(–0.687)	(–0.869)
FRESHFALL	–4.717**	–3.562
	(–2.888)	(–1.334)
GPA	12.152**	10.361**
	(9.104)	(2.927)
SATM	0.036**	0.007
	(2.689)	(0.138)
SATV	0.004	0.019
	(0.354)	(0.643)
SMALL	3.828*	3.767*
	(2.125)	(2.084)
TRANSFER	0.119	1.748
	(0.069)	(0.508)
YEARBORN	0.011	–0.412
	(0.018)	(–0.426)
λ	—	11.964
		(0.547)

*Significant at the .05 Type I error level; **significant at the .01 Type I error level.

have the advantage of a very large data set, taken from many universities. There are, however, potential problems with this type of data.

Effects of aggregation. Although aggregating the TUCE data at the class level may eliminate some idiosyncrasies of individual students, there may be detrimental effects from doing this. For example, Kennedy and Siegfried (1997, 393) conceded that

> the "no significant difference" conclusion shows up so often in so many education studies because any given technique may affect some students positively and others negatively, with no difference on the class as a whole. Our results may stem from this phenomenon: our literature review suggested that some students may be affected positively by class size and others negatively.[32]

Self-reported data. Another potential problem with the TUCE data is that much of it is self-reported by students. For example, students may have a tendency to overstate their GPAs, or otherwise be inaccurate in their reporting (Maxwell and Lopus 1994). The data for our model came from official university sources, which eliminated problems associated with self-reported data.

Bias from attrition. We have discussed the issue of attrition and the fact that large class size performance has been overstated in some TUCE studies (Becker and Powers 2001). As indicated earlier, we dropped 26 students from the sample because they withdrew from the course or did not take the final exam. The students

who withdrew had scores noticeably below the means when they withdrew, and the students who did not take the final exam received F grades. These students (23 of 26; see Table 1) were disproportionately from the large sections, as Becker and Powers (2001) would have predicted. Had we not dropped these students, our positive small class size effect would have been larger.

Measuring instructor and student performance. Aside from the potential bias problem in TUCE studies, we believe our experiment has other potential advantages. Evaluating class performance against a standardized exam (like the TUCE) is measuring two variables: student performance and instructor performance. Not only does the exam gauge how well students can retain and apply economic concepts, but it also measures the extent to which the individual instructor successfully taught the concepts tested by the TUCE exam. One would expect variation, in both student and instructor performance, across classes.

Siegfried and Kennedy (1995) provided evidence that pedagogy does not vary with class size. Nor did they find a significant difference in instructor quality among classes of different size, based on instructor self-assessments.[33] There is however, a good possibility that the material taught in different classes varies. As Saunders (1991, 269–70) noted, "less consensus exists in the profession on macro principles than on micro principles, and this poses problems for [standardized] test construction and interpretation." Specifically, this suggests that what instructors teach is important, and that, to the extent there is not a consensus on the subject matter, there is a greater likelihood for variation across classes.

With the large number of instructors taking part in collecting the TUCE data, some instructors may "teach the exam" to their students. Saunders (1991, 270) wrote,

> Before judging the adequacy of students' performance in comparison with the national norming data, however, each school and each instructor should examine TUCE III in relation to the content and purposes of their courses….Each school and each instructor must decide the extent to which the emphases of their courses agree with those of the tests.

In rare cases, perhaps the TUCE perfectly reflects the class material, in which case the instructor is unlikely to change anything when teaching the class, but some instructors, such as those whose courses do not cover all the material addressed on the TUCE, may have an incentive to cover those topics.[34] That is, even after accounting for idiosyncrasies among instructors, there may still be some instructors teaching the exam. To the extent this occurs and cannot be controlled for, it would undermine the validity of the TUCE results.

An additional concern with using the TUCE data as a measure of learning is the students' performance incentives. Although Kennedy and Siegfried (1997, 389) included a dummy for whether the TUCE exam was counted toward the students' grades (i.e., whether the student had an incentive to perform well), their measure was dichotomous rather than continuous. Different incentives may have significantly different impacts on students' effort, and this is not accounted for in studies analyzing the TUCE data.[35] We would expect that as the importance of the TUCE (in terms of students' course grades) increases, so does the incentive for the instructor to teach the exam.

A final concern, mentioned by Saunders (1991, 271), is the possibility that students could illicitly get copies of the TUCE. He suggested that security issues should be carefully considered prior to using the TUCE to affect course grades, especially if the tests are given routinely.

Using tests written specifically for the class and having the same instructor teach all classes, as we did, serves to reduce or eliminate variations that may not be adequately controlled for in the TUCE analyses.[36] We believe this to be one of the advantages of our experiment and analysis.

Potential Advantages of Using Total Points rather than Letter Grade

Watts and Lynch (1989, 237) justified using class letter grade as a measure of learning because "factors that affect learning are, in most cases, also expected to affect student grades." Raimondo, Esposito, and Gershenberg (1990, 372) cited Watts and Lynch to justify their use of course grade, "which not only measured learning in a course but also reflected the student's stock of knowledge and aptitude/ability." Hancock (1996) also used course letter grade as the dependent variable in explaining students' performance. The obvious advantage of using course letter grade instead of a standardized test score as a dependent variable is that the course grade reflects learning the material taught in the course.

However, there are at least two potential problems with using the course letter grade as the measure of student achievement or learning. First, in many studies, it is unclear exactly how student letter grades are determined. Grading scales may vary by instructor. For any given grading scale, exam scores will count, of course, but other factors may be influential. For example, if attendance, participation, homeworks, extra credit projects, or even instructor sympathy can affect the student's assigned letter grade, then this may not be an effective measure of learning.

Using total exam points accumulated, as we did, is arguably a better measure of student performance than course letter grade for another important reason. Watts and Lynch (1989, 237) acknowledged a potential "lumpiness" problem with using letter grade, rather than point totals as the measure of performance. Their specific concern was that "+" and "–" grades are not recorded at their university. Even more serious, though, is the problem of using A, B, C, D, and F grades, when, for example, an A grade is given to all students scoring 90 to 100 percent. We mitigate this lumpiness problem by using point totals, rather than letter grades, as the measure of student performance.[37] We believe this is an advantage of our analysis compared to studies (e.g., Hancock 1996) that use course letter grade as the dependent variable.

The discussion in this section is meant only to suggest that there are caveats attached to each class-size study. Large samples of standardized test data have specific advantages and limitations, as do smaller samples developed through controlled experiments, like ours.

CONCLUSION

We found a significant negative relationship between economics survey class size and student performance, as measured by exam point totals. Our positive

small class size effect would have been even stronger had we not dropped students who withdrew or did not take the final exam. Despite the fact that our experiment is limited to a single university, we believe our experiment and the results make a significant contribution to the economics class-size literature.

We attempted to hold everything constant except class size, as did Hancock (1996). Three factors may best explain why we find that "size matters" and Hancock did not. First, Hancock used letter grade rather than point totals as the measure of students' performance. Grade lumpiness may have affected Hancock's results. Second, Hancock did not account for other variables that may affect student performance. In contrast, our experiment controled for student demographic variables such as GPA and SAT scores. Third, Hancock's experiment was on statistics classes and not economics survey classes. Smaller classes and interaction with professors may be particularly well-suited to helping students to better understand the economic way of thinking.

When Becker and Powers (2001) adjusted TUCE analyses for attrition, they found a positive small class size effect on student performance. Our results are consistent with this. Still, additional research is needed. If other controlled experiments confirm our results, we would suggest the following policy implication for departments that do not have the resources to have all "small" economics principles sections. Instead of having the principles students divided into equal medium-size sections, it may be better to have one large section and numerous small sections. In this case, at least some of the students get the benefit of small classes.[38] An additional avenue for research would be the effectiveness of coupling large lecture sections with small discussion sections. This is an important issue because this is the current practice at many large universities.

Even if researchers reach a consensus that class size matters, other questions remain. Why does class size make a difference in learning economics? Even with no difference among instructors in large and small classes, perhaps students try harder in smaller classes because the professors are more likely to know their students' names (Siegfried and Kennedy 1995, 347, note 1). Students are probably more likely to attend regularly if their instructor knows them. Finally, the students may simply feel more comfortable asking and answering questions in small classes. Although the instructor in our experiment delivered the same lecture to both class sizes and allowed equal amounts of time for student questions, the dynamics of the lectures may have differed because of the different responses, attitudes and participation of the students in the different sections. Each of these effects suggests that students in smaller classes will perform better. Future experiments along these lines should include measures of student participation, such as attendance and the number of questions asked. This may shed light on why students in small classes do better even when holding pedagogy constant.[39]

The real effect of small class size probably comes through students' work ethic (including attitude toward the subject, attentiveness, shyness in class, attendance, etc.) in response to many factors, like class size, instructor attitude and personality, and so forth. These are factors over which researchers have the least control, which makes understanding them more complicated.

NOTES

1. Annual university rankings are published by a number of magazines. One of the criteria they usually report relates to class size (e.g., average class size, student-teacher ratio, or percentage of classes over a certain size).
2. Outside of economics, it appears that "size matters" is more acceptable. Details of the more general class-size debate, dating back to the 1920s, have been ably chronicled by McKeachie (1990), Mulder (1990), Slavin (1990) and Toth and Montagna (2002). Kennedy and Siegfried (1996) discuss the differences in research "cultures" among disciplines. Our discussion is confined to university survey of economics classes.
3. Kennedy and Siegfried (1996) responded to Lippman, showing the Glass and Smith (1979) framework relied upon by Lippman is flawed, as are his conclusions. Other theoretical treatments of class size include Preece (1987) and Correa (1993).
4. Bogan (1996) discussed strategies for making large classes more enjoyable and effective for students. In reviewing literature on students' attitudes, Kennedy and Siegfried (1997, 387) stated that no strong evidence exists to suggest that students are less happy in larger classes. However, students may have a different work ethic in different size classes.
5. Only students who had completed both the pretest and posttest were included in their sample (Kennedy and Siegfried 1997, 388).
6. In the extreme, of course, class size would make a difference. Slavin (1990, 7) claimed, "only by considering tutoring as a 'class size' of one is there a glimmer of truth to the conclusion that class size can significantly affect achievement."
7. Siegfried and Kennedy (1995, 350) explained the incentives to assign higher quality teachers to larger classes, though they found no evidence that this actually occurs.
8. As mentioned, Hancock's experiment is similar to ours, but his empirical analysis is quite different from ours. These differences are discussed later.
9. These include Auburn, Louisiana State, and Texas A&M.
10. Approximately 800 students enroll in economics and society each year; class size averages 50 students. Business and economics majors take a micro and macro principles sequence instead of economics and society.
11 Watts and Lynch (1989) examined the factors affecting student performance, but class size was not one of the factors they considered.
12. The advantage of using point totals instead of course grades is discussed later.
13. Course grades are mostly determined by exam performance, which ensures a strong incentive for students to perform well on the exams. Performance incentives for the TUCE are discussed later.
14. Contrary to Raimondo, Esposito, and Gershenberg's (1990) claim that multiple-choice questions "test recall and/or recognition of information," the exam questions we used required students to perform a significant amount of analysis. Few, if any, of the questions simply asked students to define or identify terms. Walstad (1998, 288) wrote, "personal preference aside, there is no established evidence to suggest that multiple-choice tests are less effective ways to measure student achievement in economics." Chan and Kennedy (2002) tested whether multiple-choice questions are easier than constructed response questions. They found little difference for questions that were "equivalent," but for more difficult questions, they found the multiple-choice questions to be easier for students. This may be part of the explanation why students prefer multiple-choice to constructed-response questions. In any case, a debate on testing formats is not the focus of this study.
15. This does not mean that it is impossible that students would share information, but even if students did discuss the exam, they did not have copies of it. For empirical analyses of cheating in economics courses, see Bunn, Caudill, and Gropper (1992), Kerkvliet (1994), and Mixon (1996).
16. For H_0: $\sigma^2_{f1}/\sigma^2_{s1} = 1$, the F statistic = 1.20 (p value = 0.333); for H_0: $\mu_{f1} - \mu_{s1} = 0$, the t statistic = 1.25 (p value = 0.216).
17. For H_0: $\sigma^2_{f0}/\sigma^2_{s0} = 1$, the F statistic = 1.01 (p value = 0.475); for H_0: $\mu_{f0} - \mu_{s0} = 0$, the t statistic = 1.55 (p value = 0.045).
18. We reject the hypothesis that the larger classes' mean is greater (t statistic = 3.60; critical value = 2.36; p value = 0.0002).
19. The ACT scores were converted to SAT equivalents using the College Board's conversion table, available at http://www.collegeboard.com/sat/cbsenior/html/stat00f.html.
20. Dropping these students did not cause the bias problems identified by Becker and Powers (2001) because the university's lack of SAT data on these students was not related to class size or to the students' class performance. In any case, 14 of these 15 students were in the large class sections; their average course grade was about 4 percent lower than the overall average. Because we

dropped these students from the model, there was probably an overstatement of the large class performance.
21. Watts and Lynch (1989, 237) also expected worse performance from freshmen.
22. If separate dummies for freshman and fall term are included instead of *FRESHFALL*, the results do not change significantly.
23. The lack of significance on *YEARBORN* is not surprising; most students enrolled in these classes are 18 or 19 years old, so there is not much variation.
24. We thank the section editor and referees for raising this issue.
25. Students were not rewarded or penalized for completing or not completing the survey.
26. We are reporting relative ratings rather than the raw ones to avoid problems of interpersonal comparisons. Readers interested in the survey details may contact the authors. One possible explanation for the students' lack of concern for class size is that the university's maximum class size is usually 50 students. Unlike large state schools, we do not have large lecture sections with hundreds of students.
27. Even if there had been a positive relationship between the factors, the direction of the relationship would be unclear. It could be that better students tend to select smaller class sizes, or that students who happen to choose smaller sections end up with higher GPAs.
28. John Jackson provided helpful advice on this section.
29. Although use of the two-step method has been criticized (in favor of a simultaneous MLE technique) when selection bias is present, the two-step method is appropriate as a test for the presence of sample selection bias (Davidson and MacKinnon 1993, 545).
30. For a discussion of probit model specification, see Olsen (1980, 1818).
31. This amounts to failing to reject the null hypothesis of no selection bias (Davidson and MacKinnon 1993, 544; Kennedy 2003, 297).
32. Kennedy and Siegfried cited Hansen, Kelly, and Weisbrod (1970) in discussing this issue.
33. Again, there are potential problems with self-reported data.
34. We do not claim to know whether such incentives exist, what they may be, or how strong they may be. But as an example, suppose some department chairs use students' TUCE performance to evaluate faculty effectiveness. In this case, there may be a strong incentive to teach the exam to students.
35. As an example, one instructor may use the TUCE score as the entire course grade. Compared to another class in which simply completing the TUCE exam counts 1 percent of the final grade, we would expect the first student to be more serious about the exam. A dummy variable cannot distinguish among different levels of performance incentive.
36. One possible disadvantage of our exams, however, is that our test questions cannot be tested as thoroughly as the TUCE questions have been.
37. Bellante (1972) also used point total, rather than letter grade, as his measure of learning.
38. This statement obviously depends on certain assumptions. For example, Siegfried and Walstad (1998, 156) suggested that at class sizes of above 20 or so, further increasing class size will not change students' behavior and performance on standardized exams.
39. The benefits of small class size may be even larger when pedagogy is adjusted for class size—if the small class instructor openly encourages questions, comments, and interaction.

REFERENCES

Becker, W. E., and J. R. Powers. 2001. Student performance, attrition, and class size given missing student data. *Economics of Education Review* 20 (4): 377–88.

Becker, W. E., and W. B. Walstad. 1990. Data loss from pretest to posttest as a sample selection problem. *Review of Economics and Statistics* 72 (1): 184–88.

Bellante, D. M. 1972. A summary report on student performance in mass lecture classes of economics. *Journal of Economic Education* 4 (1): 53–54.

Bogan, E. C. 1996. Challenges in teaching large economics classes. *International Advances in Economic Research* 2 (1): 58–63.

Bunn, D. N., S. B. Caudill, and D. M. Gropper. 1992. Crime in the classroom: An economic analysis of undergraduate student cheating behavior. *Journal of Economic Education* 23 (3): 197–207.

Chan, N., and P. E. Kennedy. 2002. Are multiple-choice exams easier for economics students? A comparison of multiple-choice and "equivalent" constructed-response exam questions. *Southern Economic Journal* 68 (4): 957–71.

Correa, H. 1993. An economic analysis of class size and achievement in education. *Education Economics* 1 (2): 129–35.

Davidson, R., and J. G. MacKinnon. 1993. *Estimation and inference in econometrics*. New York: Oxford University Press.

Finegan, T. A., and J. J. Siegfried. 1998. Do introductory economics students learn more if their instructor has a Ph.D.? *The American Economist* 42 (2): 34–46.
Glass, G. V., and M. L. Smith. 1979. Meta-analysis of research on class size and achievement. *Education Evaluation and Policy Analysis* 1 (1): 2–16.
Greene, W. H. 2003. *Econometric analysis*. 5th ed. Upper Saddle River, NJ: Prentice-Hall.
Hancock, T. M. 1996. Effects of class size on college student achievement. *College Student Journal* 30 (4): 479–81.
Hansen, L., A. Kelly, and B. Weisbrod. 1970. Economic efficiency and the distribution of benefits from college instruction. *American Economic Review.* Papers and Proceedings 60 (2): 364–69.
Heckman, J. J. 1979. Sample selection bias as a specification error. *Econometrica* 47 (1): 153–61.
Kennedy, P. E. 2003. *A guide to econometrics*. 5th ed. Cambridge, MA: MIT Press.
Kennedy, P. E., and J. J. Siegfried. 1996. On the optimality of unequal class sizes. *Economics Letters* 50 (3): 299–304.
———. 1997. Class size and achievement in introductory economics: Evidence from the TUCE III data. *Economics of Education Review* 16 (4): 385–94.
Kerkvliet, J. 1994. Cheating by economics students: A comparison of survey results. *Journal of Economic Education* 25 (2): 121–33.
Laband, D. N., and M. J. Piette. 1995. Does who teaches principles of economics matter? *American Economic Review.* Papers and Proceedings 85 (2): 335–38.
Lippman, S. A. 1990. On the optimality of equal class sizes. *Economics Letters* 33 (2): 193–96.
Lopus, J. S., and N. L. Maxwell. 1995. Teaching tools: Should we teach microeconomic principles before macroeconomic principles? *Economic Inquiry* 33 (2): 336–50.
Maddala, G. S. 1983. *Limited dependent and qualitative variables in econometrics*. New York: Cambridge University Press.
Maxwell, N. L., and J. S. Lopus. 1994. The Lake Wobegon effect in student self-reported data. *American Economic Review*. Papers and Proceedings 84 (2): 201–05.
McKeachie, W. J. 1990. Research on college teaching: The historical background. *Journal of Educational Psychology* 82 (2): 189–200.
Mixon, F. G. 1996. Crime in the classroom: An extension. *Journal of Economic Education* 27 (3): 195–200.
Mulder, J. 1990. Class size revisited: The Glass/ERS debate. *Contemporary Education* 62 (1): 47–49.
Olsen, R. J. 1980. A least squares correction for selectivity bias. *Econometrica* 48 (7): 1815–20.
Preece, P. F. W. 1987. Class size and learning: A theoretical model. *Journal of Educational Research* 80 (6): 377–79.
Raimondo, H. J., L. Esposito, and I. Gershenberg. 1990. Introductory class size and student performance in intermediate theory courses. *Journal of Economic Education* 21 (4): 369–81.
Saunders, P. 1991. The third edition of the Test of Understanding in College Economics. *Journal of Economic Education* 22 (3): 255–72.
Siegfried, J. J., and P. E. Kennedy. 1995. Does pedagogy vary with class size in introductory economics? *American Economic Review*. Papers and Proceedings 82 (2): 347–51.
Siegfried, J. J., and W. B. Walstad. 1998. Research on teaching college economics. In W. B. Walstad and P. Saunders, eds., *Teaching undergraduate economics: A handbook for instructors*, 141–66. New York: Irwin/McGraw-Hill.
Slavin, R. 1990. Class size and student achievement: Is smaller better? *Contemporary Education* 62 (1): 6–12.
Toth, L. S., and L. G. Montagna. 2002. Class size and achievement in higher education: A summary of current research. *College Student Journal* 36 (2): 253–60.
Walstad, W. B. 1998. Multiple choice tests for the economics course. In W. B. Walstad and P. Saunders, eds., *Teaching undergraduate economics: A handbook for instructors*, 287–304. New York: Irwin/McGraw-Hill.
Watts, M., and G. J. Lynch. 1989. The principles courses revisited. *American Economic Review*. Papers and Proceedings 79 (2): 236–41.