Empirical Methods in Applied Economics

Jörn-Steffen Pischke
LSE

October 2007

1 The Evaluation Problem: Introduction and Randomized Experiments

The most credible and influential research designs use random assignment. A case in point is the Perry preschool project, a 1962 randomized trial of an early-intervention program involving 123 Black preschoolers in Ypsilanti (Michigan), about half of whom were randomly assigned to an intensive intervention that included preschool education and home visits. This study is important to labor economists, since we believe that anything that affects education probably also affects earnings. It's hard to exaggerate the impact of the small but well-designed Perry experiment, which generated follow-up data through 1993 when the participants were aged 27. Dozens of academic studies cite or use the Perry findings (see, e.g., Barnett, 1992). Most importantly, the Perry project provided the intellectual basis for the massive Head Start pre-school program, begun in 1964, which ultimately served (and continues to serve) millions of American children.

1.1 The Selection Problem

We take a brief time-out for a more formal discussion of the role experiments play in uncovering causal effects. Suppose you are interested in a causal "if-then" question. To be concrete, consider a simple example: Do hospitals make people healthier? For our purposes, this question is allegorical, but it is surprisingly close to the sort of causal question health economists care about. To make this question more realistic, you might imagine we're studying a poor elderly population that uses hospital emergency rooms for primary care. Some of these patients are admitted to the hospital. This sort of care is expensive, crowds hospital facilities, and is, perhaps, not very effective (see, e.g., . Grumbach, Keane, and Bindman, 1993). In fact, exposure to other sick patients by those who are themselves vulnerable might have a net negative impact on health.

Since those admitted to the hospital get many services that are likely to be of value, on balance, the answer to the hospital-effectiveness question still seems likely to be yes. But how do we really know this? The natural approach to the empirical minded person is probably to get some data on hospital visits and compare the health status of those who have been to the hospital and those who have not. The National Health Interview Survey (NHIS) contains both types of information. It contains a question "During the past 12 months, was the respondent a patient in a hospital overnight?" which we can use to identify recent hospital visitors. It also asks "Would you say your health in general is excellent, very good, good, fair, poor?" The following table displays the mean health status (assigning a 1 to excellent health and a 5 to poor health) among those who have been hospitalized and those who have not (tabulated from the 2005 NHIS):

Group	Sample Size	Mean health status	Std. Error
Hospital	7774	2.79	0.014
No Hospital	90049	2.07	0.003

The difference in the means is 0.71, and the t-statistic for this difference is 58.9.

Taken at face value, this result suggests that going to the hospital makes people sicker, and the effect is strongly significant. Its not impossible this is the right answer: hospitals are full of other sick people who might infect us, and sharp instruments that might hurt us. Still, it's easy to see why this comparison should not be taken at face value: people who go to the hospital are less healthy to begin with. Moreover, even after hospitalization they are not as healthy as those individuals who never get hospitalized in the first place, though they may well be better than they otherwise would have been.

To describe this problem more precisely, think about hospital treatment as described by a binary variable, $\text{D}_i = \{0, 1\}$. Extending the framework to multivalued or continuous treatments is possible but the basic insights are the same. The outcome of interest, a measure of health status, is denoted by Y_i. The question is whether Y_i is *affected* by hospital care. To address this question, we must be able to imagine what might have happened to someone who went to the hospital if they had not gone and vice versa. Hence, for any individual there are two potential health variables:

$$\textit{potential outcome} = \begin{cases} \mathrm{Y}_{1i} & \text{if } \mathrm{D}_i = 1 \\ \mathrm{Y}_{0i} & \text{if } \mathrm{D}_i = 0 \end{cases} .$$

In other words, Y_{0i} is the health status of an individual had he not gone to the hospital, irrespective of whether he actually went, while Y_{1i} is the individual's health status if he goes. We would like to know the difference between Y_{1i} and Y_{0i}, which can be said to be the causal effect of going to the hospital for individual i. This is what we would measure if we could go back in time and change a person's treatment status. But because we never see both potential outcomes for any one person, we must learn about the effects of hospitalization by comparing the health of those who were and were not hospitalized.[1]

A naive comparison by hospitalization status tells us something about potential outcomes, though not necessarily what we want to know. The observed outcome, Y_i, can be written in terms of potential outcomes as

$$\begin{aligned} \mathrm{Y}_i &= \begin{cases} \mathrm{Y}_{1i} & \text{if } \mathrm{D}_i = 1 \\ \mathrm{Y}_{0i} & \text{if } \mathrm{D}_i = 0 \end{cases} \\ &= \mathrm{Y}_{0i} + (\mathrm{Y}_{1i} - \mathrm{Y}_{0i})\mathrm{D}_i. \qquad (1) \end{aligned}$$

This notation is useful because $\mathrm{Y}_{1i} - \mathrm{Y}_{0i}$ is the causal effect of hospitalization for an individual. In general, there is likely to be a distribution of both Y_{1i} and Y_{0i} in the population, so the treatment effect can be different for different individuals.

The comparison of reported health conditional on hospitalization status is formally linked to the average causal effect of hospitalization by the equation below:

$$\underbrace{E[\mathrm{Y}_i|\mathrm{D}_i = 1] - E[\mathrm{Y}_i|\mathrm{D}_i = 0]}_{\text{Observed difference in health}} = \underbrace{E[\mathrm{Y}_{1i}|\mathrm{D}_i = 1] - E[\mathrm{Y}_{0i}|\mathrm{D}_i = 1]}_{\text{average treatment effect on the treated}}$$
$$+ \underbrace{E[\mathrm{Y}_{0i}|\mathrm{D}_i = 1] - E[\mathrm{Y}_{0i}|\mathrm{D}_i = 0]}_{\text{selection bias}}$$

The term

$$E[\mathrm{Y}_{1i}|\mathrm{D}_i = 1] - E[\mathrm{Y}_{0i}|\mathrm{D}_i = 1] = E[\mathrm{Y}_{1i} - \mathrm{Y}_{0i}|\mathrm{D}_i = 1]$$

[1] The potential outcomes idea is a fundamental building block in modern research on causal effects. Important references developing this idea are Rubin (1974, 1977), and Holland (1986), who refers to frameworks involving potential outcomes as the Rubin Causal Model.

is the *average causal effect of hospitalization on those who were hospitalized.* This term captures the averages difference between the health of the hospitalized, $E[\text{Y}_{1i}|\text{D}_i = 1]$, and what would have happened to *them* had they not been hospitalized, $E[\text{Y}_{0i}|\text{D}_i = 1]$. The observed difference in health status however, adds to this causal effect a term we call *selection bias.* The selection bias is the difference in *counterfactual* health status in the no-hospital scenario between those who were and were not hospitalized. Because the sick are more likely than the healthy to seek treatment, selection bias makes those who were hospitalized seem worse of, even though they are probably better than they otherwise would have been. The goal of most empirical research is to overcome this sort of selection bias, and therefore to say something about the causal effect of a variable like D_i.

1.2 Experiments Solve the Selection Problem

Random assignment of D_i solves the selection problem because if D_i is randomly assigned it is independent of potential outcomes. To see this, note that independence means we can write

$$\begin{aligned} E[\text{Y}_i|\text{D}_i = 1] - E[\text{Y}_i|\text{D}_i = 0] &= E[\text{Y}_{1i}|\text{D}_i = 1] - E[\text{Y}_{0i}|\text{D}_i = 0] \\ &= E[\text{Y}_{1i}|\text{D}_i = 1] - E[\text{Y}_{0i}|\text{D}_i = 1] \end{aligned}$$

where we have used the (mean) independence assumption $E[\text{Y}_{0i}|\text{D}_i = 0] = E[\text{Y}_{0i}|\text{D}_i = 1]$ in the second line. In fact, given random assignment, this simplifies further

$$\begin{aligned} E[\text{Y}_{1i}|\text{D}_i &= 1] - E[\text{Y}_{0i}|\text{D}_i = 1] = E[\text{Y}_{1i} - \text{Y}_{0i}|\text{D}_i = 1] \\ &= E[\text{Y}_{1i} - \text{Y}_{0i}] \end{aligned}$$

so the effect of hospitalization on the hospitalized is the same as the effect of hospitalization on a randomly chosen patient. The main thing, however, is that random assignment of D_i eliminates selection bias for the group of experimental subjects. This does not mean that randomized trials are problem-free, but in principle they solve the most important problem that arises in empirical research. Randomized clinical trials are therefore usually considered the best possible approach to the study of causal effects.

How relevant is our hospitalization allegory? There are many examples were experiments reveal that things are indeed not what they seem. A recent example from medicine is the evaluation of hormone replacement therapy (HRT). This is a medical intervention that was recommend to middle-aged women to reduce menopausal symptoms. Evidence from the Nurses Health

Study, a large and ongoing non-experimental comparison of health-related behaviors among nurses who did and did not use HRT showed better health among the users. Evidence from a recently completed randomized trial, however, shows that some of the apparent benefits of HRT are an artifact due to selection bias and that there are serious side effects not previously detected in the simple comparison of users and non-users (see, e.g., Women's Health Initiative [WHI], Arch Intern Med. 2006;166:357-365). As it turns out, nurses who use HRT in the observation study were simply healthier anyway.

An iconic example from our own field of Labor economics is the evaluation of government-subsidized training programs. These are programs that provide a combination of classroom instruction and on-the-job training for various groups of disadvantaged workers such as the long-term unemployment, drug addicts, and ex-offenders. The idea is to improve the labor-market outcomes of these groups, especially their earnings. Many studies based on comparisons of participants and non-participants in training programs show that even after training the trainees earn less than plausible comparison groups (see, e.g., Ashenfelter, 1978; Ashenfelter and Card, 1985; Lalonde 1986). Here too, selection bias seems like a natural concern since the eligibility criteria for these programs targets men and women with low earnings potential. Not surprisingly, therefore, simple comparisons of program participants with non-participants often show lower earnings for the participants. In contrast, evidence from randomized evaluations of training programs generate mostly positive effects (see, e.g., Lalonde, 1986; Orr, et al, 1996).

Experiments in the social sciences are not as common as in medicine but they have been around for many decades, and they are becoming more prevalent. One area where the importance of random assignment is growing rapidly is education research (Angrist, 2004). The 2002 Education Sciences Reform Act now mandates the use of rigorous experimental or quasi-experimental research designs for all federally-funded education studies. A pioneering study in this area is the Tennessee STAR experiment designed to estimate the effects of placing primary school students in smaller classes.

Labor economists and others have a long tradition of trying to establish causal links between features of the classroom environment and children's learning, an area of investigation that Labor economists call "education production." This terminology reflects the fact that we think of features of the school environment as inputs that cost money, while the school output is student learning. A key question in research on education production is which inputs are worthwhile. One of the most expensive inputs is class size since smaller classes can only be had by hiring more teachers. It is therefore important to know whether the expense of smaller classes really has a payoff in terms of higher student achievement. The STAR experiment was meant to answer this question.

Many studies of education production using non-experimental data suggest there is little or no link between class size and student learning. School systems could therefore save a lot of money by hiring fewer teachers. This finding should not be taken at face value, however, since weaker students are often deliberately grouped into smaller classes. A randomized trial surmounts this difficulty by ensuring that we are comparing apples to apples, i.e., that the students assigned to classes of different sizes are otherwise comparable. Results from this randomized experiment point to a strong and lasting payoff to smaller classes (see Finn and Achilles, 1990, for the original study, and Krueger, 1999 for an econometric re-analysis).

The STAR experiment was unusually ambitious and therefore worth describing in some detail. It cost about 12 million US$ and was implemented for a cohort of kindergartners in 1985/86. The study ran for four years, i.e. until the kindergartners were in grade 3. Starting from grade 4, everybody was in regular sized classes again. The experiment involved about 11,600 children.

The average class size in regular Tennessee classes in 1985/86 was about 22.3. The idea of the experiment was to assign students to one of three treatments: small classes with 13-17 children, regular classes with 22-25 children, or regular classes with a full time teacher aide. Other regular size classes also had a part-time aide. Schools with at least three classes could choose to participate in the experiment. The 1985/86 entering kindergarten cohort was randomly assigned to one of the class types within these schools. In addition, teachers were randomly assigned to these classes.

The first question to ask about a randomized experiment is whether the randomization was carried out properly. In order to this, it is typical to look at the outcome before the treatment took place. Since the experiment should have no effect on test scores prior to the experimental intervention, the pre-treatment outcomes should be *balanced*, i.e. the averages should be the same in the treatment and control group. In this case, a pre-treatment outcome would be a student test score before the child entered kindergarten. Unfortunately, such tests were not conducted in the STAR experiment in spite of the cost. Hence, it is only possible to look at immutable characteristics of the children such as race and age, which should also be balanced across treatment and control groups.

Table 1 in Krueger (1999) presents these means. The three student characteristics are free lunch, race, and age of the child. Free lunch is a measure of family income, since only poor children qualify for a free school lunch. The means of these variables are similar across all three class types, and any differences are never statistically significant. Based on these characteristics there is no evidence that the random assignment was compromised.

The table also presents some information on the treatment and outcomes, the average class size, the attrition rate, and the test scores, measured here on a percentile scale. The attrition rate was lower in small kindergarten classrooms. This is potential a problem, which we ignore for now.[2] Class sizes are significantly lower in the assigned-to-be-small class rooms, which means that the experiment was successful in creating the desired variation.

Since a randomized experiment eliminates selection bias, the difference of the mean outcomes in the treatment and control group is an estimate of the causal effect of class size. In practice, the difference in means between treatment and control groups can be obtained from a regression of test scores on dummies for each treatment group, a point we expand on below. The estimated treatment-control differences, reported in **Krueger's Table 5,** show a small-class effect of about 5 to 6 percentile points. This difference is significantly different from zero, while the regular/aide effect is small and insignificant.

The STAR study, an exemplary randomized trial in the annals of social science, also highlights the logistical difficulty, long duration, and potentially high cost of randomized trials. In many cases, such trials are impractical.[3] In other cases, we would like an answer sooner rather than later. In many case, therefore, look for natural or quasi-experiments that mimic a randomized trial by changing the variable of interest while other factors are kept balanced.

[2] The difference in attrition rates in kindergarten means that the sample of students observed in the higher grades may not randomly distributed to class types anymore. Of course, this could even be true if the attrition *rates* were the same across class types, as long as the attrition *process* differs. The kindergarten results, which come before any attrition takes place, are therefore the most reliable.

[3] A number of problems affect the STAR data. Pupils who repeated or skipped a grade left the experiment. Students who entered an experimental school and grade later were added to the experiment and randomly assigned to one of the classes. One unfortunate, aspect of the experiment is that students in the regular and regular/aide classes were reassigned after the kindergarten year, possibly due to protests of the parents with children in the regular classrooms. There was also some switching of children after the kindergarten year. Despite these problems, the STAR experiment seems to have been an extremely well implemented randomized trial. Krueger's (1999) analysis suggests that none of these implementation problems affected the main conclusions of the study.

The quasi-experimental study of class size by Angrist and Lavy (1999) illustrates the sense in which non-experimental data can be analyzed in an experimental spirit. The Angrist and Lavy study relies on the fact that in Israel, class size is capped at 40. Therefore, a child in a fifth grade cohort of 40 students ends up in a class of 40 while a child in fifth grade cohort of 41 students ends up in a class only half as large because the cohort is split. Since students in cohorts of size 40 and 41 are likely to be fairly similar on other dimensions such as ability and family background, we can think of the difference between 40 and 41 students enrolled as being "as good as randomly assigned."

The Angrist-Lavy study therefore compares students in grades with enrollments above and below the class-size cutoffs to construct well-controlled estimates of the effects of a sharp change in class size. As in Tennessee STAR, the Angrist and Lavy results point to a strong link between class size and achievement. This is in marked contrast with naive analyses, also reported by Angrist and Lavy, based on simple comparisons between those enrolled in larger and smaller classes. These comparisons show students in larger classes doing better on standardized tests. The hospital allegory would therefore seem to apply here too.

2 Regression Analysis of Experiments

Regression is a useful tool for the study of causal questions, including the analysis of data from experiments. The discussion of the selection problem above can be cast in the language of a linear regression model. To fix ideas, suppose the treatment effect is the same for everyone, say $\text{Y}_{1i} - \text{Y}_{0i} = \rho$, a constant. In the case of a constant treatment effects we can rewrite eq. (1) in the form

$$\text{Y}_i = \underset{\substack{\shortparallel \\ E(\text{Y}_{0i})}}{\alpha} + \underset{\substack{\shortparallel \\ (\text{Y}_{1i} - \text{Y}_{0i})}}{\rho} \text{D}_i + \underset{\substack{\shortparallel \\ \text{Y}_{0i} - E(\text{Y}_{0i})}}{\eta_i} \tag{2}$$

Taking the conditional expectation of this equation with treatment status switched off and on gives

$$\begin{aligned} E[\text{Y}_i|\text{D}_i &= 1] = \alpha + \rho + E[\eta_i|\text{D}_i = 1] \\ E[\text{Y}_i|\text{D}_i &= 0] = \alpha + E[\eta_i|\text{D} = 0], \end{aligned}$$

so that,

$$E[\textsc{y}_i|\textsc{d}_i = 1] - E[\textsc{y}_i|\textsc{d}_i = 0] = \underbrace{\rho}_{\text{treatment effect}}$$
$$+ \quad \underbrace{E[\eta_i|\textsc{d}_i = 1] - E[\eta_i|\textsc{d}_i = 0]}_{\text{selection bias}} .$$

Thus, selection bias amounts to correlation between the "regression error term" and the regressor, $\textsc{d}_i$. Since

$$E[\eta_i|\textsc{d}_i = 1] - E[\eta_i|\textsc{d}_i = 0] = E[\textsc{y}_{0i}|\textsc{d}_i = 1] - E[\textsc{y}_{0i}|\textsc{d}_i = 0],$$

this correlation reflects the difference in (no-treatment) outcomes between those who get treated and those who don't. In the hospital allegory, those who were treated had poorer health outcomes in the no-treatment state, while in the Angrist and Lavy (1999) study, students in smaller classes tend to have intrinsically lower test scores.

In the STAR experiment, where $\textsc{d}_i$ is randomly assigned, the selection term disappears, and a regression of $\textsc{y}_i$ on $\textsc{d}_i$ estimates the causal effect of interest, ρ.[4] The remainder of Table 5 from Krueger (1999) shows different regression specifications, some of which include covariates other than the random assignment indicator, $\textsc{d}_i$. Covariates play two roles in regression analysis of experimental data. First, the STAR experimental design used conditional random assignment. In particular, assignment to classes of different sizes was random within schools, but not across schools. As it turns out, students attending schools of different types (say, urban versus rural) were a bit more or less likely to be assigned to a small class. Comparing all students across Tennessee, as in column (1) of Table 5, might therefore be contaminated by differences in achievement in schools of different types. Many randomized trials use conditional random assignment. To adjust for this, some of Krueger's regression models include school fixed effects, i.e., a separate intercept for each school in the STAR data. In practice, the consequences of adjusting for school fixed effects is rather minor, but we wouldn't know this without taking a look. We will more to say about regression models with fixed effects later.

[4] A detail in Krueger's analysis is that there are actually two treatment groups, one assigned to small classes and the other assigned a teacher's aide. For the sake of simplicity, and because the aide group turns out to be no different than the control group, we pretend here that there is a single treatment group only, indicated by a "small class" dummy.

The other controls in Krueger's table describe student characteristics such as race, age, and free lunch status. We saw before that these individual characteristics are balanced across class types, i.e. they are not systematically related to the class assignment of the student. If these controls, call them X_i, are uncorrelated with the treatment D_i, then they will not affect the estimate of β. In other words, estimates of ρ in the long regression,

$$\mathrm{Y}_i = \alpha + \rho \mathrm{D}_i + \mathrm{X}_i' \gamma + \eta_i' \tag{3}$$

will be close to estimates of ρ in the short regression, (2).

Nevertheless, inclusion of the variables X_i may generate more precise estimates of the causal effect of interest. Notice that the standard error of the estimated treatment effects in column (3) is smaller than the corresponding standard error in column (2). Although the controls X_i are uncorrelated with D_i, they have substantial explanatory power for Y_i. Including these controls therefore reduces the residual variance, which in turn lowers the estimated standard error. Similarly, the standard error of the estimate of ρ declines after the inclusion of the school fixed effects because these too explain an important part of the variance in student performance. The last column adds teacher characteristics. Because teachers were randomly assigned to classes, and teacher characteristics appear to have little to do with student achievement in these data, both the estimated effect of small classes and it's standard error are unchanged by the addition of the teacher variables.

TABLE I

Comparison of Mean Characteristics of Treatments and Controls: Unadjusted Data

Variable	Small	Regular	Regular/Aide	Joint *P*-Value[a]
A. Students who entered STAR in kindergarten[b]				
1. Free lunch[c]	.47	.48	.50	.09
2. White/Asian	.68	.67	.66	.26
3. Age in 1985	5.44	5.43	5.42	.32
4. Attrition rate[d]	.49	.52	.53	.02
5. Class size in kindergarten	15.1	22.4	22.8	.00
6. Percentile score in kindergarten	54.7	49.9	50.0	.00
B. Students who entered STAR in first grade				
1. Free lunch	.59	.62	.61	.52
2. White/Asian	.62	.56	.64	.00
3. Age in 1985	5.78	5.86	5.88	.03
4. Attrition rate	.53	.51	.47	.07
5. Class size in first grade	15.9	22.7	23.5	.00
6. Percentile score in first grade	49.2	42.6	47.7	.00
C. Students who entered STAR in second grade				
1. Free lunch	.66	.63	.66	.60
2. White/Asian	.53	.54	.44	.00
3. Age in 1985	5.94	6.00	6.03	.66
4. Attrition rate	.37	.34	.35	.58
5. Class size in third grade	15.5	23.7	23.6	.01
6. Percentile score in second grade	46.4	45.3	41.7	.01
D. Students who entered STAR in third grade				
1. Free lunch	.60	.64	.69	.04
2. White/Asian	.66	.57	.55	.00
3. Age in 1985	5.95	5.92	5.99	.39
4. Attrition rate	NA	NA	NA	NA
5. Class size in third grade	16.0	24.1	24.4	.01
6. Percentile score in third grade	47.6	44.2	41.3	.01

[a] p-value is for F-test of equality of all three groups.

[b] Sample size in panel A ranges from 6299 to 6324, in panel B ranges from 2240 to 2314, in panel C ranges from 1585 to 1679, and in panel D ranges from 1202 to 1283.

[c] Free lunch pertains to the fraction receiving a free lunch in the first year they are observed in the sample (i.e., in kindergarten for panel A; in first grade in panel B; etc.) Percentile score pertains to the average percentile score on the three Stanford Achievement Tests the students took in the first year they are observed in the sample.

[d] Attrition rate is the fraction that ever exits the sample prior to completing third grade, even if they return to the sample in a subsequent year. Attrition rate is unavailable in third grade.

TABLE III
DISTRIBUTION OF CHILDREN ACROSS ACTUAL CLASS SIZES BY RANDOM ASSIGNMENT GROUP IN FIRST GRADE

Actual class size in first grade	Assignment group in first grade		
	Small	Regular	Aide
12	24	0	0
13	182	0	0
14	252	0	0
15	465	0	0
16	256	16	0
17	561	17	0
18	108	36	0
19	57	76	57
20	20	200	120
21	0	378	378
22	0	594	329
23	0	437	460
24	0	384	264
25	0	175	225
26	0	130	234
27	0	54	108
28	0	28	56
29	0	29	58
30	0	30	30
Average class size	15.7	22.7	23.4

Actual class was determined by counting the number of students in the data set with the same class identification.

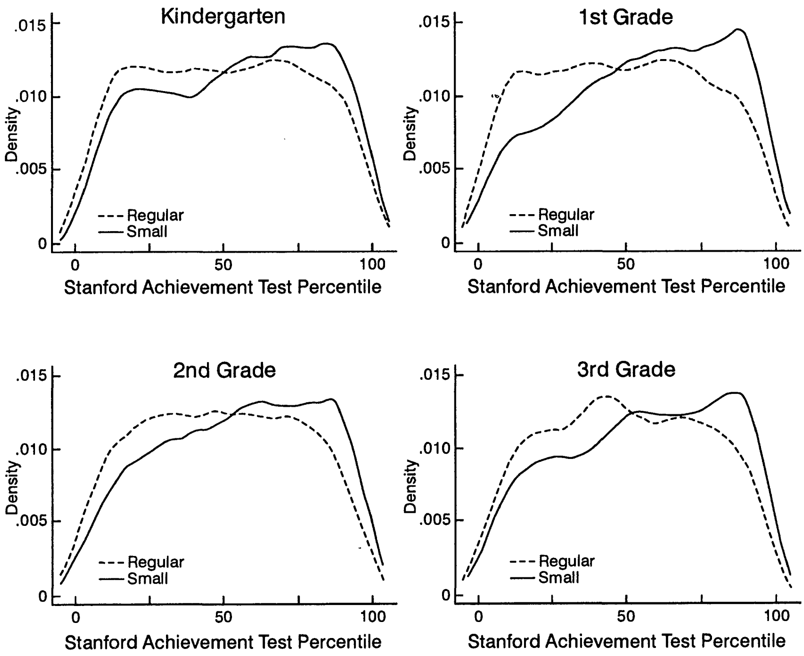

FIGURE I
Distribution of Test Percentile Scores by Class Size and Grade

TABLE V
OLS and Reduced-Form Estimates of Effect of Class-Size Assignment on Average Percentile of Stanford Achievement Test

Explanatory variable	OLS: actual class size				Reduced form: initial class size			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
A. Kindergarten								
Small class	4.82	5.37	5.36	5.37	4.82	5.37	5.36	5.37
	(2.19)	(1.26)	(1.21)	(1.19)	(2.19)	(1.25)	(1.21)	(1.19)
Regular/aide class	.12	.29	.53	.31	.12	.29	.53	.31
	(2.23)	(1.13)	(1.09)	(1.07)	(2.23)	(1.13)	(1.09)	(1.07)
White/Asian (1 = yes	—	—	8.35	8.44	—	—	8.35	8.44
			(1.35)	(1.36)			(1.35)	(1.36)
Girl (1 = yes)	—	—	4.48	4.39	—	—	4.48	4.39
			(.63)	(.63)			(.63)	(.63)
Free lunch (1 = yes)	—	—	−13.15	−13.07	—	—	−13.15	−13.07
			(.77)	(.77)			(.77)	(.77)
White teacher	—	—	—	−.57	—	—	—	−.57
				(2.10)				(2.10)
Teacher experience	—	—	—	.26	—	—	—	.26
				(.10)				(.10)
Master's degree	—	—	—	−.51	—	—	—	−.51
				(1.06)				(1.06)
School fixed effects	No	Yes	Yes	Yes	No	Yes	Yes	Yes
R^2	.01	.25	.31	.31	.01	.25	.31	.31
B. First grade								
Small class	8.57	8.43	7.91	7.40	7.54	7.17	6.79	6.37
	(1.97)	(1.21)	(1.17)	(1.18)	(1.76)	(1.14)	(1.10)	(1.11)
Regular/aide class	3.44	2.22	2.23	1.78	1.92	1.69	1.64	1.48
	(2.05)	(1.00)	(0.98)	(0.98)	(1.12)	(0.80)	(0.76)	(0.76)
White/Asian (1 = yes)	—	—	6.97	6.97	—	—	6.86	6.85
			(1.18)	(1.19)			(1.18)	(1.18)
Girl (1 = yes)	—	—	3.80	3.85	—	—	3.76	3.82
			(.56)	(.56)			(.56)	(.56)
Free lunch (1 = yes)	—	—	−13.49	−13.61	—	—	−13.65	−13.77
			(.87)	(.87)			(.88)	(.87)
White teacher	—	—	—	−4.28	—	—	—	−4.40
				(1.96)				(1.97)
Male teacher	—	—	—	11.82	—	—	—	13.06
				(3.33)				(3.38)
Teacher experience	—	—	—	.05	—	—	—	.06
				(0.06)				(.06)
Master's degree	—	—	—	.48	—	—	—	.63
				(1.07)				(1.09)
School fixed effects	No	Yes	Yes	Yes	No	Yes	Yes	Yes
R^2	.02	.24	.30	.30	.01	.23	.29	.30

TABLE V
(CONTINUED)

Explanatory variable	OLS: actual class size				Reduced form: initial class size			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
				C. Second grade				
Small class	5.93	6.33	5.83	5.79	5.31	5.52	5.27	5.26
	(1.97)	(1.29)	(1.23)	(1.23)	(1.70)	(1.16)	(1.10)	(1.10)
Regular/aide class	1.97	1.88	1.64	1.58	.47	1.44	1.16	1.18
	(2.05)	(1.10)	(1.07)	(1.06)	(1.23)	(0.87)	(0.81)	(0.81)
White/Asian (1 = yes)	—	—	6.35	6.36	—	—	6.27	6.29
			(1.20)	(1.19)			(1.21)	(1.20)
Girl (1 = yes)	—	—	3.48	3.45	—	—	3.48	3.44
			(.60)	(.60)			(.60)	(.60)
Free lunch (1 = yes)	—	—	−13.61	−13.61	—	—	−13.75	−13.77
			(.72)	(.72)			(.73)	(.73)
White teacher	—	—	—	.39	—	—	—	.43
				(1.75)				(1.76)
Male teacher	—	—	—	1.32	—	—	—	.82
				(3.96)				(4.23)
Teacher experience	—	—	—	.10	—	—	—	.10
				(.06)				(.07)
Master's degree	—	—	—	−1.06	—	—	—	−1.16
				(1.06)				(1.05)
School fixed effects	No	Yes	Yes	Yes	No	Yes	Yes	Yes
R^2	.01	.22	.28	.28	.01	.21	.28	.28
				D. Third grade				
Small class	5.32	5.58	5.01	5.00	5.51	5.42	5.30	5.24
	(1.91)	(1.22)	(1.19)	(1.19)	(1.46)	(1.08)	(1.03)	(1.04)
Regular/aide class	−.22	−.16	−.33	−.75	−.30	.12	.13	−.10
	(1.95)	(1.12)	(1.11)	(1.07)	(1.17)	(0.85)	(0.81)	(0.78)
White/Asian (1 = yes)	—	—	6.12	6.11	—	—	5.97	5.96
			(1.45)	(1.44)			(1.44)	(1.43)
Girl (1 = yes)	—	—	4.16	4.16	—	—	4.17	4.18
			(.66)	(.65)			(.66)	(.66)
Free lunch (1 = yes)	—	—	−13.02	−12.96	—	—	−13.21	−13.16
			(.81)	(.81)			(.82)	(.81)
White teacher	—	—	—	.64	—	—	—	.19
				(1.75)				(1.75)
Male teacher	—	—	—	−7.42	—	—	—	−6.83
				(2.80)				(2.76)
Teacher experience	—	—	—	.04	—	—	—	.03
				(.06)				(.06)
Master's degree	—	—	—	1.10	—	—	—	.88
				(1.15)				(1.15)
School fixed effects	No	Yes	Yes	Yes	No	Yes	Yes	Yes
R^2	.01	.17	.22	.23	.01	.16	.22	.22

All models include constants. Robust standard errors that allow for correlated residuals among students in the same class are in parentheses. Sample size is 5861 for kindergarten, 6452 for first grade, 5950 for second grade, and 6109 for third grade.