# The Central Role of Noise in Evaluating Interventions that Use Test Scores to Rank Schools

Kenneth Y. Chay
University of California, Berkeley

Patrick J. McEwan
Wellesley College

Miguel Urquiola
Columbia University

Revised: January 2005

For helpful comments we thank Michael Boozer, David Card, David Figlio, Sebastian Galiani, Hidehiko Ichimura, Robert LaLonde, Lant Pritchett, Douglas Staiger, and especially Thomas Lemieux and three anonymous referees. We are also grateful to Cristián Cox and many officials of Chile’s Ministry of Education for providing data.

# Abstract

Many programs reward or penalize schools based upon students' average performance. Kane and Staiger (2001) warn that imprecision in the measurement of school-level test scores could impede these efforts. There is little evidence, however, on how seriously noise hinders the evaluation of the impact of these interventions. We examine these issues in the context of Chile's P-900 program – a country-wide intervention in which resources were allocated based on cutoffs in schools' mean test scores. We show that transitory noise in average scores and mean reversion lead conventional estimation approaches to greatly overstate the impacts of such programs. We then show how a regression discontinuity (RD) design that utilizes the discrete nature of the selection rule can be used to control for reversion biases. While the RD analysis provides convincing evidence that the P-900 program had significant effects on test score gains in 4th grade, these effects are much smaller than is widely believed.

Every state government in the United States has begun to measure student achievement, aggregate the results, and rate the performance of public elementary and secondary schools.[1] A growing number of states further use test-based rankings to allocate rewards, assistance, and sanctions to schools.[2] Not surprisingly, there is considerable interest in the impact of such interventions on student test scores.[3]

Kane and Staiger (2001, 2002a) have noted that mean test scores may provide a noisy measure of school performance due to large error variances, particularly among smaller schools. They conclude that mean test scores from a single year can provide a misleading ranking of schools. For example, a school's appearance at the bottom (or top) of a ranking may be the result of transitory bad (or good) luck in the testing year, and may not be indicative of the school's true performance.

This paper examines an important implication of these findings. If transitory testing noise, due to luck or sampling variation, is mean reverting, conventional evaluation approaches will yield misleading estimates of the effect of interventions that use test-based rankings to select schools. For example, suppose that schools with very low mean scores in a given year are selected to receive an intervention (e.g., assistance or sanctions). If the previous reasoning is correct, then the measured poor performance of such schools is, in part, a result of having obtained a strongly negative error in the program assignment year. Unless errors are perfectly correlated over time, one would expect subsequent test scores in such schools to rise, even in the absence of the intervention. Thus, the measured test score gains from a straightforward difference-in-differences analysis will reflect a combination of a true program effect and spurious mean reversion.[4] The dilemma is similar to that observed in evaluations of training programs in

[1] Under the federal *No Child Left Behind Act of 2001*, all states must use test score results to assess whether public schools are making "adequate yearly progress" (Education Week, 2004).

[2] As of 2004, 16 states used rankings to allocate rewards to high-performing schools; 36 provided assistance to low-performing ones; and 27 administered sanctions – such as closure or the withholding of funds – to low-performing schools (Education Week, 2004). Other countries have experimented with rewarding schools or teachers based upon the test score performance of students, such as Israel (Lavy, 2002), Kenya (Glewwe, Ilias, and Kremer, 2003), Mexico (McEwan and Santibañez, 2004), and Chile, the subject of this paper.

[3] See Hanushek and Raymond (2002), and the citations therein, for a recent overview of the literature on test-based accountability. They note that the nascent literature often focuses on the "gaming" responses of school personnel to the enactment of reforms, rather than the reform's impact.

[4] For instance, Kane and Staiger (2002b) suggest that mean reversion led North Carolina officials to erroneously conclude that test score gains among low-achieving schools were due to a program of targeted assistance.

which assignment to the program is based on pre-program earnings.[5]

To date, the literature has not reached a consensus on how severe the biases introduced by mean reversion can be, or on how best to address them. We analyze this issue in the context of Chile's 900 Schools Program (hereafter referred to as P-900). Beginning in 1990, the program identified approximately 900 schools which had low mean fourth-grade test scores in 1988. In the first three years of the program – the focus of this paper – program participation was strongly determined by whether a school's mean score fell below a cutoff value in its region. Treated schools received infrastructure improvements, instructional materials, teacher training, and tutoring for low-achieving students.

We find that transitory noise in average scores, and the resulting mean reversion, lead conventional estimation approaches to greatly overstate the positive impact of P-900. For example, difference-in-differences estimates suggest that P-900 increased 1988-1992 test score gains by 0.4-0.7 standard deviations; yet using P-900-type assignment rules, we can generate similar effects during earlier periods (1984-1988) in which the program was not yet in operation. Further, schools chosen for P-900 exhibit a sharp decline in test scores just before the program year, which is consistent with a negative shock in average scores in the year used to assign program participation.

To address this problem, we implement a regression discontinuity approach that exploits the discrete relation between program selection and pre-program test scores. We also derive a simple analytical framework for measuring and eliminating the effect of the error variance in test scores on the mean reversion bias. We find that P-900 resulted in no test score gains from 1988 to 1990, the first year of its operation, but that it did increase 1988-1992 test score gains by about 0.2 standard deviations. Graphical analyses and several robustness checks provide complementary evidence that comparing the gains of schools just above and just below the assignment cutoff effectively eliminates reversion biases. Finally, the strategies illustrated herein should be applicable whenever tests or other "pre-scores" – in concert with assignment cutoffs – are used to allocate a program.

[5] Specifically, those enrolled in the training program often experienced negative labor market shocks right before enrollment, relative to those who were not enrolled. In short, Ashenfelter's "dip" is relevant whenever treatment assignment is based on noisy, pre-treatment values of the outcome variables, whether these are earnings or test scores (e.g., Angrist and Kruger, 1999; Ashenfelter, 1978; Ashenfelter and Card, 1985; Heckman, LaLonde, and Smith, 1999).

## I. Background on P-900

In 1990, Chile's government introduced the P-900 program, a package of four interventions targeted at low-performing, publicly-funded schools (García-Huidobro, 1994, 2000; García-Huidobro and Jara Bernardot, 1994).[6] First, schools received improvements in their infrastructure, such as building repairs. Second, schools were given a variety of instructional materials, including textbooks for students in grades 1 through 4, small classroom libraries, cassette recorders, and copy machines. Third, teachers in these grades attended weekly training workshops conducted by local supervisors of the Ministry of Education. The workshops were focused on improving pedagogy in the teaching of language and mathematics. Fourth, the program created after-school tutoring workshops that met twice a week, and were attended by 15 to 20 third and fourth graders who were not performing at grade level. Each workshop was guided by two trained aides recruited from graduates of local secondary schools.

The first two years of the program (1990 and 1991) focused on the provision of infrastructure and instructional materials (García-Huidobro, 2000). In 1992, the program expanded to include in-service training and after-school workshops. Several program officials, including an early administrator, indicated to us that the in-service training and afternoon workshops constituted the bulk of effort and expenditure. Further, all treated schools apparently received the interventions with roughly the same intensity. Unfortunately, there is no administrative record of what each school received, and, as such, we cannot separately identify the contribution of each component of the intervention. We therefore treat P-900 as a "black box" and estimate the combined impact of its components.

In addition to the effects of resource investments, the program may have affected schools in other ways. First, teachers and administrators might have raised their effort levels in response to the identification of their schools as poorly performing, especially given that government officials openly described the program as "intensive care" (Cox, 1997). It is also possible that they reduced effort in the

[6] "Low-performing" was simply defined as the schools obtaining the lowest mean performance, unadjusted for the characteristics of the students that they enroll. About 90 percent of enrollments in Chile are in public and private schools that receive voucher-style government subsidies. All of these institutions were eligible for P-900. "Elite" private schools, which charge tuition and do not receive public subsidies, account for the remaining 10 percent of enrollments. These were not eligible for the program. For details on Chile's system of school finance, see McEwan and Carnoy (2000) and Hsieh and Urquiola (2003).

hope of receiving additional resources from the program. Second, P-900 may have encouraged the children of some households to exit or enter the treated schools. One might expect the former if parents interpreted program selection as a signal that the institution was not adequately serving their children. The latter could result if they thought their children could benefit from additional resources.

The program's initial assignment occurred in two stages (García-Huidobro and Jara Bernardot, 1994; García-Huidobro, 2000). The first relied on achievement tests administered to the population of fourth-graders in 1988.[7] Officials of the Ministry of Education calculated each school's mean in language, mathematics, and the combination of both subjects. These scores were ordered from highest to lowest within each of Chile's 13 administrative regions. Separate cutoff scores were established for each region, and schools below their region's cutoff were pre-selected to participate. It bears emphasis that the 1988 scores were collected under a different government (before the return of democratic elections), at which time the P-900 program was not even contemplated. It is therefore not plausible that schools sought to manipulate their performance in 1988 in order to qualify for the program in 1990.

In the second stage, regional teams of officials reviewed each list and some pre-selected schools were removed from eligibility. The decisions were apparently based on two criteria. First, very small or inaccessible schools did not participate, in order to reduce program costs, and also because another program (MECE-Rural) would eventually be created for them. Second, schools were excluded if they demonstrated managerial problems – such as private voucher schools that misreported their enrollments, an offense subject to legal penalties. Finally, there is the possibility that regional teams introduced unobserved criteria for school eligibility. However, schools themselves appear to have had little scope to refuse the program, in part because all costs were covered by the national government.

In the past, P-900 has been lauded as a success, and the previous literature reflects a widespread perception that it substantially raised the achievement of treated schools.[8] The empirical basis of this

[7] The test scores were collected as part of the SIMCE (*Sistema de Medición de la Calidad de la Educación*) and included both public and private schools. In practice, some schools were excluded from the testing because of their extremely low enrollments. In total, the excluded schools accounted for no more than 10 percent of total enrollments, and they were not eligible for P-900.

[8] See, for example, García-Huidobro (1994), García-Huidobro (2000), Angell (1996), Gajardo (1999), World Bank (1999), and Tokman (2002).

perception can be easily replicated. Suppose that we observe the mean fourth-grade achievement of each school at two different points in time. We can assess whether mean achievement increases more among P-900 schools than among untreated schools using a difference-in-differences framework,

$$(1) \qquad \Delta y_j = \overline{y_j^{90}} - \overline{y_j^{88}} = \alpha + \theta \cdot P900_j + \varepsilon_j^{90} - \varepsilon_j^{88}$$

where $\overline{y_j^t}$ is the average score across fourth graders in school *j* at time *t*, and $\Delta y_j$ is the change in the mean from 1988 to 1990 (hereafter the "gain score"); $P900_j$ is a dummy variable equal to one if the school received the treatment, and $\varepsilon_j^t$ are the unobserved school level factors at time *t*. The parameter of interest is $\theta$, which measures the gain for treated schools relative to untreated schools.

Table 1 reports descriptive statistics for fourth-grade language and mathematics gain scores in 1988-1990 and 1988-1992.[9] Note that the 1988 *combined* mean scores were used to assign the program, and that this assignment remained in force through 1992. Further, 1990 was the first full year of treatment, with all tests administered at the end of the respective school years. Using these data to estimate (1) yields large and statistically significant estimates of $\theta$. For example, the 1988 to 1990 gain scores imply program effects that are equivalent to 0.30 and 0.49 standard deviations in math and language, respectively. For 1988 to 1992 gain scores, the implied P-900 effects are 0.45 and 0.68 standard deviations, respectively.[10] Thus, it is not surprising that many observers have concluded that the P-900 program has been one of the most successful schooling interventions in the developing world.

## II. Evaluation Problems Due to Testing Noise

For these estimates to have a causal interpretation, it must be the case that the differences in the gain scores of treated and untreated schools are entirely due to the program. In this section, we argue that

[9] Test score data is available after 1992, but we do not use it for three reasons. First, the program selection rules became increasingly nebulous. Some schools were removed from the treatment group and they were replaced by others. The selection of these schools apparently relied more on the subjective opinions of Ministry personnel and less on a strict assignment rule. Second, the Ministry of Education initiated a large reform of primary schools with World Bank support – the MECE program. The program started on a small scale in 1992, but rapidly expanded during the next six years to the universe of publicly-funded schools. It is not known whether the program was more or less likely to be targeted at P-900 schools. Third, we presume that schools became increasingly aware of P-900 selection rules, and may have sought to obtain low scores in order to participate, undermining the application of the regression-discontinuity design.

[10] These estimates are from the full sample of schools. The subsequent regression-discontinuity estimates will rely upon a smaller subsample of urban schools with larger enrollments.

mean reversion – the outgrowth of imprecisely measured mean test scores – causes this condition to be violated. Further, it is a plausible explanation for the large effects found in previous evaluations. We derive a simple analytical framework to illustrate the role of testing noise in mean reversion bias.

Figure 1 (Panel A) presents a stylized version of the actual assignment rule. It plots the average score of each school in the year used to determine school rankings (referred to as “pre-score”) on the x-axis, and the treatment status of the school, assuming a value of 0 or 1, on the y-axis. The pre-score ranges from 0 to 100, and we arbitrarily choose 50 as the program cutoff. That is, all schools with pre-scores of 50 or less are treated, and the rest are not.

Panel B illustrates a visual analogue of the case in which the difference-in-differences estimate may be unbiased. The y-axis and x-axis display the post-program gain score and the pre-score, respectively. Here, the vertical distance between the two line segments is the added gain among P-900 schools – i.e., the treatment effect. A causal interpretation of the effect may be justified, since pre-scores and gain scores are not otherwise related.

If, on the contrary, schools with lower pre-scores have higher gain scores even in the absence of P-900, the situation might resemble Panel C, where there is no program effect – that is, there is no break in the relation between the gain score and the pre-score close to the assignment cutoff. Nevertheless, a regression specification like (1) will erroneously suggest a positive treatment effect by fitting a difference in means without allowing for mean reversion.

Why might schools with lower pre-scores have higher gain scores? The answer is rooted in noisy measures of schools’ mean test scores (Kane and Staiger, 2001). First, there may be one-time events that influence test scores, such as a school-wide illness or distraction from construction noise in the school’s vicinity. Second, there is sampling variation in test scores, since each cohort of students that enters a school is analogous to a random draw from a local population. Thus, a school’s mean test score will vary with the specific group of students starting school in any given year. This variance, in turn, depends on two factors: the variability of performance in the population from which the school draws its students, and the number of students in the grade tested. We cannot directly assess the first of these, but we can verify the implication that scores should be more variable in schools with lower enrollments.

Figure 2 (Panel A) plots each school's average 1988 language score against its fourth grade enrollment in 1988. It reveals that mean performance is substantially more variable among smaller schools. Similarly, Panel B plots 1988 to 1990 language gain scores against combined enrollment in 1988 and 1990. It shows a strong negative relationship between school-level variation in gain scores and school enrollments.

Of critical importance is the fact that extreme scores in 1988 occur among schools with lower enrollments (e.g., less than 30 students in the fourth grade). This suggests that some schools obtained very low scores in 1988, and therefore qualified for P-900, simply because they experienced an "unlucky" circumstance or a bad draw of fourth-grade students that year. Since they are unlikely, on average, to experience a bad draw again in 1990 or 1992, their average achievement will tend to rise – i.e., they will revert towards the mean – even in the absence of the P-900 program. Thus, mean reversion poses a serious challenge to any evaluation of such a program.[11]

More formally, consider a simple model for individual, fourth-grade test scores in 1988 and 1990:

$$y_{ij}^{88} = \lambda_j + u_j^{88} + \alpha_i^{88} \tag{2}$$

$$y_{ij}^{90} = \lambda_j + u_j^{90} + \alpha_i^{90} \tag{3}$$

where *i* indexes students and *j* indexes schools; $\lambda_j$ is a school-level permanent effect on student scores; $u_j^t$ is a school-level transitory shock (e.g., construction noise) in year *t*; and $\alpha_i^t$ is student *i*'s "test-taking" ability in fourth-grade cohort *t*. We assume that $\lambda_j, u_j^t$, and $\alpha_i^t$ are independent of each other. We further assume that $\alpha_i^{88}$ and $\alpha_i^{90}$ are independent, which is plausible since students in each year are drawn from different cohorts. Suppose further that $u_j^{88}$ and $u_j^{90}$ are independent and that $\lambda_j \overset{D}{\sim} iid\,(0,\sigma_\lambda^2), u_j \overset{D}{\sim} iid\,(0,\sigma_u^2), \alpha_i \overset{D}{\sim} iid\,(0,\sigma_\alpha^2)$.[12]

[11] While noise in the treatment selection variable (due to luck or sampling variation) leads to some "randomization" in the treatment assignment, this will still result in mean reversion bias if the noise is transitory and the selection variable is equal to previous values of the outcome variable. Thus, our context for applying the regression discontinuity design differs substantially from the static, cross-sectional case in which the selection variable is not the outcome variable.

[12] The independence of $u_j^{88}$ and $u_j^{90}$ may be viewed as a strict assumption since it stipulates that school-level shocks die out within two years, which may not be true if the shock is a severe natural disaster or a regime change at the school (e.g., a new administration). In addition, the independence of the three variance components will be violated if, for example, better schools tend to attract more or different students. These possibilities do not detract

In this scenario, the average test scores at the school-level in 1988 and 1990 have the form:

$$\overline{y_j^{88}} = \lambda_j + u_j^{88} + \sum_{i\in j} \frac{1}{N_j^{88}} \alpha_i^{88} \tag{4}$$

$$\overline{y_j^{90}} = \lambda_j + u_j^{90} + \sum_{i\in j} \frac{1}{N_j^{90}} \alpha_i^{90}\ , \tag{5}$$

and it follows that the variance of the 1988 to 1990 school gain score is:

$$Var(\overline{y_j^{90}} - \overline{y_j^{88}}) = 2\sigma_u^2 + \sigma_\alpha^2 \left( \frac{N_j^{88} + N_j^{90}}{N_j^{88} N_j^{90}} \right). \tag{6}$$

Thus, the variation in gain scores across schools is a function of the variance in the transitory school-level shock ($\sigma_u^2$), the variance in individual testing abilities ($\sigma_\alpha^2$), and the enrollment sizes in the two years ($N_j^{88}$ and $N_j^{90}$). Although we do not have student-level testing data, equation (6) implies that one can estimate the student testing variance by regressing the sample variances of 1988 to 1990 gain scores among schools with the exact same number of students in 1988 and 1990 on a constant and $\left( \frac{N_j^{88} + N_j^{90}}{N_j^{88} N_j^{90}} \right)$.[13] The estimated constant (multiplied by 0.5) provides an estimate of the variance of the transitory school effect, $\sigma_u^2$, and the estimated slope coefficient provides an estimate of the variance of the student effects, $\sigma_\alpha^2$.

Implementing this regression with 1988-1990 language gain scores yields an estimate (and sampling error) of $\sigma_u^2$ equal to 14.7 (2.6) and an estimate of $\sigma_\alpha^2$ equal to 586.5 (54.8).[14] To gauge the reliability of these estimates, Figure 2 (Panel B) plots the 1st and 99th percentiles of 1988-1990 language gain scores implied by the estimates of $\sigma_u^2$ and $\sigma_\alpha^2$, further assuming that $u_j^t$ and $\alpha_i^t$ are normally distributed.[15] The implied 1-to-99 interval fits the actual school-level variation in gain scores quite well. First, it emulates the decreasing variation as total enrollment increases – particularly the "funneling"

from the expositional usefulness of the model. Further, below we estimate models that do not rely on these restrictions, with virtually no change in the estimated effects of P-900.

[13] In other words, individual-level variances have implications for the variance in average scores across schools with the same number of students in 1988 and 1990. Unfortunately, student level data are not available for these years.

[14] For 1988-1990 mathematics gain scores, the estimate of $\sigma_u^2$ is 12.6 (2.8), and the estimate of $\sigma_\alpha^2$ is 726.0 (58.9).

[15] Thus, the 99th and 1st percentiles are equal to $\mu \pm \Phi^{-1}(0.99) \cdot \sqrt{2\sigma_u^2 + \sigma_\alpha^2 \left( \frac{N_j^{88} + N_j^{90}}{N_j^{88} N_j^{90}} \right)}$, where $\Phi^{-1}$ is the inverse of the standard normal cumulative distribution function and $\mu$ is the average 1988-1990 language gain score.

effect at total enrollments below 150. In addition, only 2.4 percent of the school observations lie outside the interval, which is close to the 2 percent predicted by the normality assumption.[16]

Using this framework, one can describe how mean reversion potentially introduces bias in difference-in-differences estimates of the P-900 effects. Recall that such bias arises from a relation between 1988 test scores, which are used to determine selection into the program, and test score gains. Consider the "regression" coefficient relating the 1988-1990 gain score to the 1988 average score:

$$\rho = \frac{Cov(\overline{y_j^{90}} - \overline{y_j^{88}}, \overline{y_j^{88}})}{Var(\overline{y_j^{88}})} = \frac{Cov(\overline{y_j^{88}}, \overline{y_j^{90}})}{Var(\overline{y_j^{88}})} - 1 = \frac{\sigma_\lambda^2}{\sigma_\lambda^2 + \sigma_u^2 + \frac{\sigma_\alpha^2}{N_j^{88}}} - 1 . \tag{7}$$

This slope coefficient would prevail *even in the absence* of the P-900 intervention. It clarifies three scenarios under which mean reversion may, or may not, bias the results.

First, the slope coefficient is zero (and mean reversion is absent) if variance in test scores is entirely due to permanent differences across schools ($\sigma_\lambda^2$). If this unlikely circumstance prevails, then the difference-in-differences estimate will be unbiased (as illustrated in Figure 1, Panel B).

Second, the slope coefficient becomes negative and approaches -1 as: (1) the variance due to student heterogeneity ($\sigma_\alpha^2$) and the variance due to school-level transitory shocks ($\sigma_u^2$) increase; and (2) as the number of students enrolled in a school ($N_j^{88}$) gets small. If all schools have the same enrollment, then the (negative) slope coefficient is the same for all schools. This form of mean reversion is illustrated by the straight lines in Figure 1 (Panels C and D). However, the assumption of identical enrollments across all schools conflicts with the data.

Third, the slope coefficient will vary across schools to the extent that they enroll different numbers of students. In particular, the coefficient will be more negative for smaller schools. Figure 2 (Panel A) already illustrated that smaller schools were more likely to obtain very low average scores in 1988, and therefore more likely to qualify for the program. To further illustrate this point, Panel C (Figure 2) plots nonparametric predicted values of schools' 1988 enrollment against their 1988 average score, and shows that schools with extreme scores (particularly low ones) tend to have small enrollments.

[16] Further, 9.2 percent of the observations lie outside the implied 5-to-95 percentile interval, and 47.2 percent lie outside the implied 25-to-75 interquartile range – both close to the 10 and 50 percent implied by normality.

This is consistent with the mean reversion illustrated in Panels E and F of Figure 1. Specifically, since the schools in the tails of the 1988 test score distribution are, on average, smaller than the schools in the middle of the distribution, the mean reversion pattern will have a cubic polynomial shape with steeper negative slopes at very low and very high 1988 scores. Figure 2 (Panel D) shows the nonparametric fit of the 1988-1990 language gain score as a function of the 1988 test score selection variable (from a local linear regression smoother). The estimated conditional mean has slightly steeper slopes at very low and high scores, consistent with a cubic polynomial shape.

### III. Regression Discontinuity Approaches to Mean Reversion

The discrete nature of the program selection rule facilitates a quasi-experimental regression discontinuity (henceforth, RD) design to control for mean reversion biases.[17] We illustrate several approaches that produce consistent estimates of the treatment effect as long as the reversion bias is "smooth" at the regional test score cutoff determining selection into the P-900 intervention.

Building upon equation (1), the goal is to eliminate sources of correlation between $P900_j$ and $(\varepsilon_j^{90} - \varepsilon_j^{88})$, such as mean reversion. First, we can use equations (4) and (5) to re-write (1):

$$(8) \qquad \overline{y_j^{90}} - \overline{y_j^{88}} = \alpha + \theta \cdot P900_j + \left( u_j^{90} + \sum_{i \in j} \frac{1}{N_j^{90}} \alpha_i^{90} \right) - \left( u_j^{88} + \sum_{i \in j} \frac{1}{N_j^{88}} \alpha_i^{88} \right), \text{ and}$$

$$(9) \qquad P900_j = 1\left( \overline{y_j^{88}} < y^* \right) = 1\left( \lambda_j + u_j^{88} + \sum_{i \in j} \frac{1}{N_j^{88}} \alpha_i^{88} < y^* \right),$$

where $1(\cdot)$ is an indicator function that is equal to one if the enclosed statement is true and $y^*$ is the cutoff for P-900 eligibility in the school's region. As long as $u_j^t$ and $\alpha_i^t$ are nontrivial and transitory and school enrollments are finite, there will be a positive correlation between $P900_j$ and $(\varepsilon_j^{90} - \varepsilon_j^{88})$.

The key to addressing this problem is to note that P-900 assignment is a *discrete* function of 1988 average test scores. This implies that one can control for "smooth" functions of 1988 scores to control for mean reversion bias while still estimating the P-900 effect. As long as the chosen function of 1988 scores

[17] For a history and overview of the RD approach, see Campbell and Stanley (1963) and especially Shadish, Cook, and Campbell (2002). The RD design has recently been used to explore a range of issues in the economics of education (Angrist and Lavy, 1999; Guryan, 2002; Jacob and Lefgren, 2004a, 2004b; Kane, 2003; van der Klaauw, 2002; Urquiola, 2000).

absorbs the reversion bias, the resulting estimates will be consistent. The previous discussion suggests three candidates for control functions. The first is simply a linear term in 1988 average scores:

$$(10) \qquad \Delta y_j = \alpha + \theta \cdot P900_j + \beta_1 \overline{y_j^{88}} + \Delta\varepsilon_j ,$$

which is sufficient if the mean reversion function is linear, as illustrated in Panels C and D of Figure 1.

Second, we anticipate that the mean reversion function may have a cubic polynomial shape, since smaller schools are more likely to have scores in the tails of the test score distribution. Thus, one can directly adjust for a cubic polynomial in 1988 average scores:

$$(12) \qquad \Delta y_j = \alpha + \theta \cdot P900_j + \beta_1 \overline{y_j^{88}} + \beta_2 \overline{y_j^{88}}^2 + \beta_3 \overline{y_j^{88}}^3 + \Delta\varepsilon_j .$$

This specification will control for mean reversion patterns such as those in Panels E and F of Figure 1.

A third approach is to adjust for the mean reversion pattern implied by the model:

$$(11) \qquad \Delta y_j = \alpha + \theta \cdot P900_j + \left( \frac{\sigma_\lambda^2}{\sigma_\lambda^2 + \sigma_u^2 + \frac{\sigma_\alpha^2}{N_j^{88}}} - 1 \right) \cdot \overline{y_j^{88}} + \Delta\varepsilon_j .$$

This allows the intensity of mean reversion to vary with schools' enrollments. It can be estimated via nonlinear least squares.[18] It is possible that the "control function" in (11) is misspecified in its arguments, $N_j^{88}$ and $\overline{y_j^{88}}$, due to violations of the underlying assumptions of the model. Thus, we also estimate specifications in which $N_j^{88}$, $\overline{y_j^{88}}$, and their interaction enter flexibly into the regression model to gauge the robustness of the findings from estimating equation (11).

A stricter approach is to estimate the regression for the subsample of schools within arbitrarily narrow bands close to the cutoff point, *y** (e.g., Angist and Lavy, 1999, van der Klaauw, 2002). If other factors affecting gain scores are similar for schools just above and below the cutoff, then comparing the gain scores in treated and untreated schools with pre-scores close to the cutoffs will control for all omitted factors correlated with being selected for P-900, including the intensity of mean reversion. Under this

[18] Note that equation (11) cannot be used to estimate the student-level variance ($\sigma_\alpha^2$); thus, we restrict its value to the estimate obtained from equation (6) when we apply nonlinear least squares. We also restrict the school-level transitory variance ($\sigma_u^2$) to be equal to the estimates obtained from equation (6), although this is not necessary. In less restrictive regression results, not reported here, the estimates of $\sigma_u^2$ are similar to those obtained from (6).

assumption, discrete differences in mean gain scores between treated and untreated schools close to the cutoff can be attributed to P-900. In Figure 1, Panels D and F depict a stylized version of this situation, where the treatment effect is identified as the break in the relation between the gain and the pre-score close to the discontinuity. Below, we also adjust for a cubic polynomial in 1988 scores even when focusing on narrow subsamples of schools.

Before proceeding, two further empirical challenges must be addressed: selection and sorting. The selection issue arises because the program's initial allocation – based on a strict assignment rule – was occasionally subverted when administrators removed some schools from program participation.[19] This raises the possibility that assignment is correlated with unobserved determinants of achievement. We carry out three exercises to address this issue.

First, we repeat the previous analysis while excluding Region 13 (the "Metropolitan Region," composed primarily of Santiago) from the sample. This is the region in which administrators seem to have exercised the most discretion, by far.[20]

Second, we apply an instrumental variables approach that uses the indicator function, $1\left(\overline{y_j^{88}} < y*\right)$, as an instrument for P-900 status. Even if $P900_j$ is correlated with $\Delta\varepsilon_j$ because of selection, one can still obtain consistent estimates of $\theta$ by instrumenting $P900_j$ with the indicator for initial program eligibility (equal to one if a school falls below the regional cutoff). In this case, the first stage and reduced-form equations of the two-stage least squares estimator are:

(13) $P900_j = \delta_0 + \delta_1 \cdot ELIGIBLE_j + \delta_2 \overline{y_j^{88}} + \delta_3 \overline{y_j^{88}}^2 + \delta_4 \overline{y_j^{88}}^3 + v_j$, and

(14) $\Delta y_j = \pi_0 + \pi_1 \cdot ELIGIBLE_j + \pi_2 \overline{y_j^{88}} + \pi_3 \overline{y_j^{88}}^2 + \pi_4 \overline{y_j^{88}}^3 + \Delta\eta_j$, where

$ELIGIBLE_j = 1\left(\overline{y_j^{88}} < y*\right) = 1\left(\lambda_j + u_j^{88} + \sum_{i \in j} \frac{1}{N_j^{88}} \alpha_i^{88} < y*\right)$, and

$\delta_1$ provides an estimate of the discrete jump in the probability of treatment for schools below the cutoff;

[19] Thus, the P-900 program is closer to the "fuzzy" RD design that is described by Shadish et al. (2002).

[20] In previous work, we excluded a larger subset of regions, including Region 13, and were left with a sample in which the selection rule correctly assigned at least 95% of schools to their treated or untreated status (Chay, McEwan, and Urquiola, 2003). This yields a similar pattern of results. In fact, the Metropolitan Region, by virtue of its size, accounts for the vast majority of mis-assigned schools (i.e., those where administrators exercised discretion by ignoring the initial selection).

$\pi_1$ is the discrete difference in test gain scores between schools below and above the cutoff. The instrumental variable estimate is equal to $\pi_1 / \delta_1$.

Third, we apply an entirely different identification strategy, facilitated by the variation in cutoff scores across Chile's 13 regions.[21] Instead of comparing treated and untreated schools close to cutoffs within regions, we compare treated and untreated schools across regions with the same pre-scores.

In addition to selection, sorting poses a problem because it is possible that families responded to P-900 by withdrawing or enrolling their children in treated schools, potentially altering the distribution of observed and unobserved student attributes across treated and untreated schools. A straightforward way of addressing sorting (as well as selection) is to include controls for schools' observable socioeconomic status (SES) in the above specifications.

## IV. Program Assignment

The first stage of program assignment relied on the combined mean of 1988 fourth-grade test scores, in concert with assignment cutoffs that were specific to each of Chile's 13 regions. To illustrate this, Figure 3 reports data for Region 9, in the south of Chile. In Panel A, each dot represents a school, ordered on the basis of its 1988 average score (on the x-axis). On the y-axis, a one indicates a school that received the P-900 treatment, while a zero indicates untreated schools. The figure highlights two important features of program assignment.

First, assignment did not rely exclusively on 1988 test scores because there are treated and untreated schools with similar 1988 test scores, particularly in the left side of the distribution. This is consistent with official accounts of the assignment process that were discussed in a previous section. Regional teams from the Ministry of Education excluded some "pre-selected" schools, particularly if these were small and rural. In light of this, Panel B restricts attention to urban schools with 15 or more students in the fourth grade (henceforth referred to as urban, larger schools).[22] There is evidently a

[21] Tyler, Murnane, and Willett (2000) use a similar approach in their analysis of the GED.

[22] Varying the 15 student threshold somewhat does not affect the conclusions reached below.

discrete change in the probability of treatment in Region 9 among these schools, providing a good setting for an RD analysis. Of course, this judgment must be made separately for each of Chile's 13 regions.

Second, the exact regional cutoffs are not observed. Panel A assumes that the cutoff is located at the rounded-up integer of the highest 1988 test score observed among treated schools (56 in Region 9). We refer to this as "cutoff definition 1." This definition is accurate if administrators only *remove* schools from the treatment that fall below the initial cutoff, consistent with published accounts. However, Panel B makes it clear that administrators may have also *added* schools to the treatment with 1988 test scores above the cutoff. In the case of Region 9 (see Panel B), there is one such school in the subsample of urban, larger schools. In such a case, "cutoff definition 1" will yield cutoffs that are too high.

To further explore this issue, Table 2 summarizes the data from each of Chile's 13 regions.[23] Columns 1 and 2 of the table present sample sizes – both the total and that which remains after restricting attention to urban, larger schools. Column 3 contains "cutoff definition 1." Columns 4 and 5 present the percentage of schools that are classified correctly using this definition – i.e., a school with an average 1988 score below (above) the cutoff actually received (did not receive) the treatment. As expected, the cutoffs perform better in the urban, larger school sample, where four regions have correct classification rates above 90 percent. In contrast, Region 13 has, by far, the worst rate of correct assignment (only 35 percent) and conditioning on urban, larger schools has little effect on this rate.[24] In addition, nationwide, only 59 percent of the schools are correctly classified.

The visual evidence from Region 9 (and other regions) suggests that the true cutoff falls below a few outlying high scores among the schools that receive P-900. The issue can be formalized as the problem of determining the break point for a regression discontinuity design when the true break point is not known. In another paper (Chay, McEwan, and Urquiola, 2005), we propose an approach for determining the break points in such data. The approach sets the break for P-900 eligibility at the score that maximizes the goodness-of-fit from a model of P-900 participation as a function of an indicator equal

[23] With the exception of Region 13 (comprised mostly of Santiago), the regions' numbers correspond to a north-to-south geographic order. Thus, Region 1 is the northernmost area of Chile, Region 12 is the southernmost area, and Region 13 is located just south of Region 5.

[24] This is to be expected since Region 13, composed mainly of Metropolitan Santiago, contains very few smaller or rural institutions.

to one if the school's score is below a particular threshold.[25] This simplifies to choosing as the cutoff the 1988 average score that maximizes the estimated difference in the probability of P-900 selection between schools just below and above the cutoff.

These scores are shown in column 6, and we refer to them as "cutoff definition 2."[26] The results for this second definition are presented in columns 7 and 8. In the urban, larger school sample, at least 90 percent of schools are classified correctly in 9 regions, and this figure is 85 percent or more in the remaining 4 regions. Nationwide, approximately 92 percent of the urban, larger schools are correctly classified as being in or out of the P-900 program. As mentioned above, administrators in Region 13 exercised the most discretion in removing "eligible" schools from the program. When this region is excluded from the sample, "cutoff definition 2" correctly classifies the P-900 status of *95 percent* of all urban, larger schools in Chile.

Figure 3 (Panel C) describes the result of this exercise for the nationwide sample of urban, larger schools. In order to pool the data across regions, we create a variable that indicates each school's score relative to its respective regional cutoff. This simplifies the presentation of the results and will eventually facilitate the estimation of an average, nationwide P-900 effect. The figure plots unweighted smoothed values of the proportion of schools treated, with respect to their distance from their respective regional cutoff score. As expected, there are sharp changes in the probability of treatment close to the cutoff, an essential component of the RD approach. Panel C also plots fitted values of a regression of P-900 on eligibility (essentially the "first stage" described in equation (13), but without additional controls). It shows that probability of treatment is 0.67 higher among eligible schools.

Finally, Panel D presents the same information, but excluding Region 13 from the sample. As expected, the changes in the probability of treatment are even more pronounced, and the probability of treatment is 0.82 higher among eligible schools.

[25] We thank an anonymous referee for suggesting this exercise. Also see Kane (2003).

[26] In an earlier draft of this paper (Chay, McEwan, and Urquiola, 2003) we used the rounded-up integer of the 95 percentile score among P-900 schools in each region as cutoff definition 2. In Chay, McEwan, and Urquiola (2005), we find that the "optimal" regional cutoffs are the previous cutoffs minus 0.6 points. Using the previous definition leads to very similar findings to those presented in this paper. In addition, we get similar findings when we allow the optimal break points to be different distances from the 95th percentile score in different clusters of regions (see Chay, McEwan, and Urquiola, 2005).

## V. Results

A simple difference-in-differences analysis suggests that P-900 had a substantial effect on fourth-grade gain scores. Columns 1 and 6 in Table 3 illustrate this for math and language, respectively. The coefficients on the treatment dummy are always statistically significant and substantial. For 1988-1990 gain scores, the effects are equivalent to 0.23 and 0.45 standard deviations of the math and language test score distributions, respectively. For 1988-1992 gain scores, the effect sizes are equal to 0.39 and 0.66 standard deviations of the respective distribution.

### A. Evidence on mean reversion

If test scores are indeed a noisy measure of performance, however, then a portion of these estimates is likely due to mean reversion. Further, if this is the case we should find "P-900 effects" even in periods in which no program existed. To verify this, we draw on test scores collected in 1984.[27] As a first exercise, we identified a sample of 1,546 schools with scores available in 1984 and 1988. To maintain comparability with other estimates, we restrict the sample to include urban schools with 15 or more students in the fourth grade. We then ranked schools according to their 1984 average score and, roughly simulating the actual P-900 selection rule, designated the lowest 20 percent as "treated." Of course, P-900 did not exist in this period, and there were no similar schemes. Unless driven by mean reversion, the fictitious treatment should yield no effect. In fact, estimating (1) with 1984-1988 math gain scores yields an estimate for $\theta$ of 4.0, somewhat larger than that for 1988-1992. This suggests that mean reversion is indeed a primary concern in evaluating this type of program.

To provide additional time series evidence on this issue, Figure 4 uses 1,534 schools with test scores in 1984, 1988, 1990, and 1992 (again, limiting the sample to urban, larger schools). Panels A and B show the annual mean score of P-900 and non-P-900 schools, respectively.[28] The key observation is that scores for treated schools display a "dip" in 1988. A plausible interpretation is that many schools

[27] These test scores were collected under a different system, the PER (*Programa de Evaluación de Rendimiento*), and were applied to a smaller sample of schools. No test scores were collected in the years between 1984 and 1988.

[28] Test scores within each year are standardized to a mean of 50 and a standard deviation of 10.

experienced transitory negative shocks in 1988, leading them to be selected. By 1990, mean reversion returned their scores close to their 1984 levels. Importantly, the opposite story can be told of Panel B, where untreated schools experience a slight upward "bounce" in 1988. This is consistent with positive shocks that are followed by mean reversion. Nonetheless, the bounce is less pronounced in Panel B because the untreated schools are drawn from a less extreme part of the 1988 test score distribution.

In short, mean reversion appears to pose a substantial challenge to the evaluation of programs like P-900. The remainder of the paper addresses this challenge with an RD design. It relies upon the expectation that we should observe fewer fluctuations like those in Panels A and B (Figure 4) among schools close to regional cutoff scores. Panel C illustrates this by presenting the mean difference in test scores between P-900 and untreated schools for three sets of schools: all schools (the bottom line) and those within 5 and 2 points of their respective regional cutoffs (the lines in the middle and at the top of the figure, respectively).

In 1984, the difference between treated and untreated schools in the full sample was equal to about 10 points. In 1988, the year of assignment, it increased to almost 15. By 1990 however, the difference was again almost exactly equal to 10 points. However, the raw differences are smaller when the sample is restricted to schools that are close to their regional cutoffs, and the dips are much less pronounced. This is consistent with this difference being less influenced by unusually high or low scores that noise would induce in the extremes of the distribution. It suggests that an RD approach can be a valuable way of addressing the problem of mean reversion.

Finally, we note that Figure 4 foreshadows some results in the next section. Panels A and C suggest that treated schools experienced transitory shocks in 1988, and that by 1990 they had returned to their previous performance. In both cases, however, slight improvements are visible by 1992, implying that the program may have had a real effect on achievement.

### B. Regression discontinuity results

Panels A through D in Figure 5 plot unweighted smoothed values of schools' gain scores against their 1988 pre-scores (relative to their respective regional cutoff), distinguishing between P-900 and untreated schools. There is a negative relation between gain scores and 1988 scores, which is consistent with substantial mean reversion.[29] Further, the pattern of mean reversion is reminiscent of Figure 1 (Panel E) in that it is generally more intense for schools well below the cutoff – although the same does not hold for those with extremely high scores. This partially reflects the fact that the RD sample does not include schools with extremely low fourth grade enrollments – i.e., below 15 students.

To the extent that P-900 had an effect, we should observe a break in these relationships *close to the cutoff* (one analogous to that in Figure 1, panels D and F). The graphs for 1988-1990 gain scores (Figure 5, Panels A and B) suggest no such break – the P-900 and non-P-900 lines essentially overlap at the cutoff. Nevertheless, a "naive" evaluation would suggest that P-900 had a large effect in its first year. Panels C and D, which refer to 1988-1992 gain scores, present a different picture. Here a break is visible and is equal to roughly 2 points.[30]

The regression results are consistent with the visual evidence. Table 3 adds increasingly flexible specifications of the 1988 average score to control for mean reversion, as well as SES controls and regional fixed effects.[31] When columns 2 and 7 include controls for schools' 1988 score (relative to their regional cutoff), the P-900 coefficients fall substantially, particularly for 1988-1990 gain scores (in which case they are essentially zero). Columns 3 and 8 estimate equation (11) via nonlinear least squares, restricting the student-level variance to be equal to the estimates from the previous section. The coefficient estimates for the P-900 indicator are comparable. These specifications also yield estimates of $\sigma_{\lambda}^{2}$, which we return to in the conclusion.

[29] In both periods the average gains are substantial, and few schools had negative gain scores. This is consistent with anecdotal evidence indicating that the tests became somewhat less difficult over time. It could also be evidence that schools are "teaching to the test."

[30] In fact, two types of breaks are visible in these figures: those at the cutoff and those for treated and untreated schools with overlapping 1988 scores (our initial regression evidence will capture both), and their magnitude is generally similar. This reflects the fact that the assignment discontinuities (Figure 3) are imperfect, leading to "fuzziness" across the cutoff score. In view of this, the most unrestricted graphical representation of the effects is obtained by calculating smoothed values separately for the P-900 and untreated schools, as we do in Figure 5.

[31] In other regressions, not reported, we also include flexible specifications of the 1988 language or mathematics score, corresponding to the dependent variable. These did not yield substantively different results. For a subset of these schools, we can further include a polynomial of 1984 test scores. We omit these specifications because they reduce the sample size while not substantively affecting any of the conclusions below.

The next two specifications include a cubic in the 1988 score, with some small changes in the P-900 coefficients. Finally, columns 5 and 10 attempt to control for selection and sorting by including controls for SES, as well as regional dummies. This leads to slight increases in the P-900 coefficients (less than half a point). Overall, the P-900 coefficients' magnitude is consistent with no effect for 1988-1990 gain scores and an effect of about 2 points for 1988-1992 gain scores.[32] This conclusion holds even when we limit the sample to schools that fall within increasingly narrow bands near the cutoff point for each region. Table 4 presents the results from such regressions for the 1988-1992 gain scores. (We omit similar estimates for 1988-1990 gains since they simply reaffirm the finding of no effect.) Here again there are statistically significant effects close to two points, for both language and mathematics, for the subsamples of schools within 5, 3, and 2 points of the regional cutoffs.

To summarize, we find no evidence that P-900 had produced a positive effect by 1990, but we do find effects on 1988-1992 gain scores of approximately two points, roughly equal to 0.2 standard deviations. What could account for the lack of an effect on 1988-1990 gain scores? The 1990 cohort of fourth-graders in P-900 schools participated for a single year (the 1990 test was administered towards the end of the school year). Thus, one possibility is that a single year of exposure was insufficient to affect achievement. Another possibility is that the program – as it was implemented – was different in 1990 (recall that fewer program elements were available in the first year). Regarding the positive effects in 1988-1992 gain scores, one interpretation is that scores increase by a small amount for each year of exposure to P-900 (i.e., 0.07 standard deviations).[33] Another is that the effect was obtained in a single grade, but this cannot be determined from the data.

[32] As discussed above, we also estimated models that allow 1988 average scores and school enrollments to enter more flexibly into the regression specification than in equation (11). Specifically, we included a cubic in scores, a quadratic in enrollments, and the interaction of scores with enrollments (we also estimated more flexible specifications with no substantive change in the results). Table A.1., available in the online appendix, presents the results. Columns 2 and 6 show that, as predicted by the model, school enrollments are a significant predictor of test score gains conditional on 1988 scores. The other columns show that the estimates of the P-900 effect are not sensitive to more flexible specifications of the "control function". Further, models that include school sizes fit the data better than the model that only includes a cubic in 1988 test scores – i.e., have higher adjusted $R^2$'s.

[33] One might also ask whether test score gains are justified by program costs. An early cost analysis of the program estimated an annualized cost of 26 dollars per student in 1992, which is 11% of the typical school-wide cost per student (Peirano R. and McMeekin, 1994). Whether the program is cost-effective can only be determined by obtaining cost-effectiveness ratios for other programs, information that is not available.

### C. Robustness

For a first robustness check, we restricted the sample by excluding Region 13, the one with the most evidence of administrative discretion in P-900 selection. Panels E and F in Figure 5 present the graphical evidence for this subsample, with similar breaks evident in the vicinity of the cutoff. The regression results (available in Table A.2 in the online appendix) are statistically significant and generally close to those in Tables 3 and 4, despite the reduced sample sizes.

As a second check, we instrument for P-900 status using the indicator of program eligibility in equations (13) and (14). Table 5 (Panel A) presents the first-stage regression results. As Figure 3 (Panels C and D) had already suggested, eligible schools – i.e., those falling below the cutoff – have a substantially higher probability of being treated. This finding is robust to the inclusion of a cubic in 1988 scores as well as other controls. In the subsamples that exclude Region 13 (columns 5 and 6) the eligibility coefficient is even larger. All regressions explain a large proportion of the variance in treatment status.

Panels B and C contain the estimates of the reduced-form model, in which math and language gains are regressed on the indicator of P-900 eligibility. The eligibility coefficient is analogous to the intent-to-treat estimator in a randomized experiment. The coefficients are always positive and significant at conventional levels. The simple difference-in-differences estimates for the impact of eligibility in column 1 are similar in magnitude to the difference-in-differences estimates of the P-900 effects in Table 3. This highlights that even if one has a valid instrument for a program like P-900 – e.g., the case where pure "noise" causes schools to be under or over the cutoffs – mean reversion may still bias the instrumental variables estimates of the program effects on test score gains.[34]

The instrumental variables (IV) estimates, reported in panels D and E of Table 5, are simply the ratio of the reduced-form coefficients to the corresponding first-stage coefficients. They provide

[34] This point is illustrated in Panels A and B of Figure A.1 (in the online appendix) for the 1988-1992 gain scores. The dotted lines plot the fitted values of a regression of the average gain score on only an eligibility dummy (drawn from column 1 in Table 5). The solid lines plot those from a regression on the same dummy *and* the 1988 score (relative to the regional cutoff), thereby introducing the simplest control for mean reversion (see column 2 in Table 5). The estimated break at the cutoff is substantially smaller in the latter case. Panels C and D in Figure A.1 plot nonparametric predictions separately for the samples of eligible and ineligible schools. They show a clear break in 1988-1992 test score gains for eligible versus ineligible schools near the regional cutoffs.

estimates of the effect of the treatment on the treated. They are, naturally, somewhat larger than the reduced-form estimates. They are also similar in magnitude to the OLS estimates of the program effects shown in the preceding tables, suggesting that unobservable selection may not be an important source of bias.[35]

For a third robustness check, the data permit an entirely different identification strategy, facilitated by regional variation in cutoffs. Instead of comparing treated and untreated schools *within* regions on either side of pre-score cutoffs, we can compare treated and untreated schools with similar pre-scores *across* regions. As an example, consider the sample of schools with mean 1988 scores that are greater than 49.4 and less than or equal to 51.4. In Regions 1, 3, 4, 11, and 12, these schools are below the corresponding regional cutoff (and subject to the treatment). In contrast, schools in regions 2, 5-10, and 13 are above the corresponding regional cutoff (and not subject to the treatment). Hence, they can serve as a counterfactual. This assumes that the effect of the treatment does not vary across regions and that the choice of cutoff across regions is exogenous.[36]

Table A.4 (available in the online appendix) focuses on 1988-1992 test score gains and summarizes the results for five feasible "experiments," each conducted within successive ranges of 1988 scores. The point estimates are more variable, but are generally positive and consistent with the estimates found earlier (with the exception of Panel B). In fact, estimates from the pooled sample that controls for

[35] The results in Table 3 suggest that selection and sorting (on observed SES) lead to small downward biases in estimates of the program effect. To examine this further, Table A.3 (available in the online appendix) report results from regressions of the 1990 and 1992 SES indices on the indicator variables for P-900 treatment and for P-900 eligibility for various samples. For the full sample of schools, column 1 in Panels A and B shows that treated schools have much lower average SES than untreated schools. Although the difference is greatly reduced by controlling for a cubic in 1988 average scores (i.e., the control function) and by focusing on the subsamples of schools near the regional cutoffs (columns 2-6), it is still statistically significant. While the findings imply a small role for sorting (e.g., the coefficients on the P-900 indicator are similar for the 1990 and 1992 SES regressions), they are consistent with the possibility that the schools scoring below the regional cutoff that were removed from the P-900 program list had higher average SES than the eligible schools that were not removed. Columns 7-12 in Panels A and B report the results from the same exercise in a sample that excludes Region 13. The estimated P-900 coefficient is much smaller and statistically insignificant after controlling for a cubic in 1988 scores. Panels C and D report the results from specifications in which the dependent variable is the indicator for program eligibility (i.e., whether a school falls below the regional cutoff). In all samples, there is no relationship between eligibility and SES after controlling for a cubic in 1988 scores. This suggests that the estimates in Table A.3 and in Table 5 may be purged of the potential selection biases in the OLS estimates in Table 3.

[36] We have little direct evidence on how regional cutoffs were chosen. For example, they do not simply reflect the fact that the 20th percentile within each region varies; indeed, different proportions of schools appear to have been treated across regions.

interval dummy variables in Panel F yield results quite close to those observed above (about 2 points in gain scores). In addition, the estimates are robust to the inclusion of additional controls for 1988 test scores, suggesting that limiting the sample to narrow ranges of initial test scores removes a substantial amount of the reversion bias.

As a final check on mean reversion, we employed 1988 and 1992 data to create two fictitious programs: "P-450" and "Reverse P-900".[37] The first is obtained by selecting the bottom 10 percent of schools, as opposed to the bottom 20 percent that was roughly applied in the case of P-900. "Reverse P-900," in turn, selects the *top* 20 percent of performers in 1988. The expectation, given our claims about mean reversion, is that in simple difference-in-differences specifications, "P-450" should produce estimates larger than those for P-900, while "Reverse P-900" should yield estimates of a negative sign, and perhaps similar in magnitude. Further, in regressions that control for a cubic in 1988 scores and use subsamples of schools within narrow bands around the fictitious cutoffs, these effects should be diminished or disappear entirely. Table A.5 in the online appendix confirms these expectations, using the full sample of schools.[38]

## VI. Conclusion

In the perpetual search for policies to improve educational quality, many governments have turned to interventions that use test-based school rankings to allocate resources, rewards, or sanctions. Not surprisingly, there is a growing demand for knowledge on the effect of these interventions. This paper has shown that noise and the consequent mean reversion produce important complications in the evaluation of such schemes. The use of intuitively-appealing evaluation strategies, like difference-in-differences, can lead to dramatically biased estimates of the program effects.

[37] We are indebted to an anonymous referee and seminar participants for suggesting this analysis.

[38] The simple difference-in-differences specifications in column 1 suggest strongly positive effects for P-450 (slightly larger than a corresponding estimate of 7.34 for P-900 in the same sample). The effect for "Reverse P-900" are comparable in magnitude but of the opposite sign. Columns 2-3 repeat these analyses with additional controls and within narrow bands. The "program" effects disappear and are statistically insignificant, as one would anticipate. For reasons of space, we present results for 1988-1992 language scores. Those from other subject/year combinations yield similar conclusions.

That is certainly the case in previous evaluations of Chile's P-900 program. Our results suggest that a regression discontinuity methodology can potentially circumvent this problem. In the case of P-900, it reveals that the program's effects, while positive, are much smaller than the previous estimates. Just as importantly, the issues our findings raise are germane to the ongoing evaluation of similar programs – including many states' educational accountability reforms – that assign treatments on the basis of high or low pre-score measures. To the extent that assignment is based on strict cutoffs, the methods used in this paper are a useful means of addressing potential biases.

The results also suggest that noise might limit the ability of student testing to identify "bad" or "good" schools. Employing the estimates of $\sigma_u^2$, $\sigma_\alpha^2$, and $\sigma_\lambda^2$ in this paper, it is straightforward to calculate the percentage of variance in 1988 scores that is due to transitory rather than permanent components. For a school with the median fourth-grade enrollment in Chile of 30, 33 and 21 percent of the variance in language and math scores, respectively, is due to transitory testing noise. For a school with the 25th percentile enrollment of 15, the corresponding percentages are higher (44 and 30 percent). Thus, many smaller schools were eligible for the P-900 program due to transitory negative shocks to their average test scores in 1988. This finding is especially relevant for the P-900 program since it had a compensatory intent: the government desired to improve the lowest achieving schools, thereby aiding low-income children who presumably disproportionately populate such institutions. Our results suggest that the achievement of this goal was hampered by school-level sampling variation in test scores.

## References

Angell, A. (1996). Improving the quality and equity of education in Chile: The Programa 900 Escuelas and the MECE-Basica. In A. Silva (Ed.), *Implementing policy innovations in Latin America: Politics, economics, and techniques* (pp. 94-117). Washington, DC: Inter-American Development Bank.

Angrist, J.D., and Krueger, A.B. (1999). Empirical strategies in Labor Economics. In O. Ashenfelter & D. Card (Eds.), Handbook of Labor Economics (Vol. 3A, pp. 1277-1366). Amsterdam: Elsevier.

Angrist, J. D., & Lavy, V. (1999). Using Maimonides' rule to estimate the effect of class size on scholastic achievement. *Quarterly Journal of Economics, 114* (2), 533-575.

Ashenfelter, O. (1978). Estimating the effect of training programs on earnings. *Review of Economics and Statistics, 60*(1), 47-57.

Ashenfelter, 0., & Card, D. (1985). Using the longitudinal structure of earnings to estimate the effect of training programs. *Review of Economics and Statistics, 67*(3), 658-660.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research.* Boston: Houghton Mifflin.

Chay, K. Y., McEwan, P. J., & Urquiola, M. (2003). *The central role of noise in evaluating interventions that use test scores to rank schools.* Working Paper No. 10118, National Bureau of Economic Research, Cambridge, MA.

Chay, K., McEwan, P. J., & Urquiola, M. (2005). *Implementing the regression discontinuity design when the break point is unknown: Evidence from a schooling intervention in Chile.* In progress, University of California, Berkeley.

Cox, C. (1997). *La reforma de la educación chilena: Contexto, contenidos, implementación.* Documentos de Trabajo 8, PREAL, Santiago.

Education Week. (2004). State of the states. *Education Week, 23*(17), 97-99.

Gajardo, M. (1999). *Reformas educativas en América Latina: Balance de una década.* Documentos de Trabajo 15, PREAL, Santiago.

Garcia-Huidobro, J. E. (1994). Positive discrimination in education: Its justification and a Chilean example. *International Review of Education, 40*(3-5), 209-221.

Garcia-Huidobro, J. E. (2000). Educational policies and equity in chile. In F. Reimers (Ed.), *Unequal schools, unequal chances: The challenges to equal opportunity in the Americas* (pp. 161-178). Cambridge, MA: Harvard University, David Rockefeller Center for Latin American Studies.

Garcia-Huidobro, J. E., & Jara Bernardot, C. (1994). El Programa de las 900 Escuelas. In M. Gajardo (Ed.), *Cooperación internacional y desarrollo de la educación* (pp. 39-72). Santiago: Agencia de Cooperación Internacional.

Glewwe, P., Ilias, N., & Kremer, M. (2003). *Teacher incentives.* Working Paper No. 9671, National Bureau of Economic Research, Cambridge, MA.

Guryan, J. (2002). *Does money matter? Regression-discontinuity estimates from education finance reform in Massachusetts.* Working Paper No. 8269, National Bureau of Economic Research, Cambridge, MA.

Hanushek, E. A., & Raymond, M. E. (2002). Improving educational quality: How best to evaluate our schools? In Y. K. Kodrzycki (Ed.), *Education in the 21st century* (pp. 193-236). Boston: Federal Reserve Bank of Boston.

Heckman, J. J., LaLonde, R. J., & Smith, J. A. (1999). The economics and econometrics of active labor market programs. In O. Ashenfelter & D. Card (Eds.), *Handbook of labor economics* (Vol. 3) (pp. 1865-2097). Amsterdam: Elsevier.

Hsieh, C.-T., & Urquiola, M. (2003). *When schools compete, how do they compete'? An assessment of Chile's nationwide school voucher program.* Working Paper No. 10008, National Bureau of Economic Research, Cambridge, MA.

Jacob, B. A., & Lefgren, L. (2004a). The impact of teacher training on student achievement: Quasi-experimental evidence from school reform efforts in Chicago. *Journal of Human Resources*, 39(1), 50-79.

Jacob, B. A., & Lefgren, L. (2004b). Remedial education and student achievement: A regression-discontinuity approach. *Review of Economics and Statistics*, 86(1), 226-244.

Kane, T. J. (2003). *A quasi-experimental estimate of the impact of financial aid on college- going.* Working Paper No. 9703, National Bureau of Economic Research, Cambridge, MA.

Kane, T. J., & Staiger, D. O. (2001). *Improving school accountability measures.* Working Paper No. 8156, National Bureau of Economic Research, Cambridge, MA.

Kane, T. J., & Staiger, D. O. (2002a). The promise and the pitfalls of using imprecise school accountability measures. *Journal of Economic Perspectives*, 16(4), 91-114.

Kane, T. J., & Staiger, D. O. (2002b). Volatility in test scores: Implications for test-based accountability systems. In D. Ravitch (Ed.), *Brookings Papers on Education Policy* (pp. 235-283). Washington, DC: Brookings Institution Press.

Lavy, V. (2002). Evaluating the effect of teachers' group performance incentives on pupil achievement. *Journal of Political Economy, 110(6),* 1286-1317.

McEwan, P. J., & Carnoy, M. (2000). The effectiveness and efficiency of private schools in Chile's voucher system. *Educational Evaluation and Policy Analysis*, 22(3), 213-239.

McEwan, P. J., & Santibañez, L. (2004). *Teacher incentives and student achievement: Evidence from a large-scale reform.* Unpublished manuscript, Wellesley College and RAND.

Peirano R., C., & McMeekin, R. W. (1994). Gastos y costos del Programa de las 900 Escuelas en el período enero 1990-junio 1993. In M. Gajardo (Ed.), *Cooperación internacional y desarrollo de la educación* (pp. 73-97). Santiago: Agencia de Cooperación Internacional.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi- experimental designs for generalized causal inference.* Boston: Houghton Mifflin.

Tokman, A. (2002). *Evaluation of the P900 program: A targeted education program for underperforming schools.* Documento de Trabajo No. 170, Banco Central de Chile, Santiago.

Tyler, J. H., Murnane, R. J., & Willett, J. B. (2000). Estimating the labor market signaling value of the GED. *Quarterly Journal of Economics, 115*(2), 431-468.

Urquiola, M. (2000). *Identifying class size effects in developing countries: Evidence from rural schools in Bolivia. The Review of Economics and Statistics,* forthcoming.

van der Klaauw, W. (2002). Estimating the effect of financial aid offers on college enrollment: A regression-discontinuity approach. .*International Economic Review,* 43(4), 1249- 1287.

World Bank. (1999). *Educational change in Latin America and the Caribbean: A World Bank strategy paper.* Washington, DC: World Bank.

## Table 1: Descriptive Statistics

| | 1988-1990 sample | | | 1988-1992 sample | | |
|---|---|---|---|---|---|---|
| | All | Non P-900 | P-900 | All | Non P-900 | P-900 |
| Math score, 1988 | 48.9 | 50.9 | 40.6 | 49.8 | 52.4 | 40.6 |
| | (10.2) | (10.1) | (5.0) | (9.7) | (9.1) | (5.0) |
| Language score, 1988 | 50.7 | 53.0 | 41.0 | 52.1 | 55.2 | 41.1 |
| | (11.9) | (11.8) | (6.5) | (11.4) | (10.6) | (6.4) |
| Math gain score, 1988-1990 | 6.0 | 5.4 | 8.4 | | | |
| | (9.9) | (10.1) | (8.9) | | | |
| Language gain score, 1988-1990 | 5.0 | 4.1 | 8.8 | | | |
| | (9.5) | (9.4) | (8.9) | | | |
| Math gain score, 1988-1992 | | | | 13.4 | 12.4 | 16.7 |
| | | | | (9.6) | (9.2) | (10.0) |
| Language gain score, 1988-1992 | | | | 11.4 | 10.1 | 16.2 |
| | | | | (9.0) | (8.5) | (9.1) |
| P-900 | 0.19 | | | 0.22 | | |
| Urban | 0.59 | 0.62 | 0.50 | 0.69 | 0.74 | 0.51 |
| 4th-grade enrollment (median) | 30 | 31 | 29 | 38 | 41 | 29 |
| [25%-tile, 75%-tile] | [15, 63] | [13, 68] | [20, 48] | [21, 70] | [21, 78] | [21, 49] |
| SES index, 1990 | 54.8 | 57.7 | 42.7 | 59.7 | 64.4 | 42.9 |
| | (29.5) | (30.1) | (23.3) | (27.8) | (27.2) | (22.9) |
| SES index, 1992 | | | | 44.3 | 49.4 | 26.3 |
| | | | | (30.3) | (30.0) | (23.9) |
| Sample size | 4,628 | 3,741 | 887 | 3,878 | 3,016 | 862 |

Notes: Standard deviations are in parentheses. Test scores are expressed as the percentage of items correct. *P-900* is a dummy variable indicating program treatment. *Urban* is a dummy variable indicating urban (versus rural) location. *4th-grade enrollment* reports the number of fourth-graders who took the SIMCE test in 1988, and whose scores comprise the school-level average. The *SES index* measures student socioeconomic status (SES), as reported by JUNAEB (*Junta Nacional de Auxilio Escolar y Becas*). It is scaled 0-100, with higher values indicating higher SES.

Table 2: P-900 cutoff definitions, sample sizes, and percentage correctly classified by region

| | Number of Schools | | Cut-off definition 1 | | | Cut-off definition 2 | | |
|---|---|---|---|---|---|---|---|---|
| | | | Cut-off Score | % correctly classified: | | Cut-off Score | % correctly classified: | |
| Region | All schools | Urban larger schools | | All schools | Urban larger schools | | All schools | Urban larger schools |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| 1 | 65 | 49 | 52 | 89.2 | 100.0 | 51.4 | 87.7 | 98.0 |
| 2 | 76 | 70 | 50 | 84.2 | 91.4 | 49.4 | 84.2 | 91.4 |
| 3 | 59 | 47 | 52 | 83.1 | 95.7 | 51.4 | 83.1 | 95.7 |
| 4 | 266 | 95 | 55 | 53.0 | 87.4 | 51.4 | 61.3 | 94.7 |
| 5 | 493 | 333 | 53 | 64.7 | 74.5 | 49.4 | 78.7 | 89.8 |
| 6 | 373 | 124 | 49 | 55.0 | 75.8 | 43.4 | 81.8 | 97.6 |
| 7 | 494 | 157 | 48 | 57.5 | 79.6 | 42.4 | 77.5 | 97.5 |
| 8 | 768 | 377 | 52 | 53.6 | 69.5 | 43.4 | 77.3 | 97.1 |
| 9 | 359 | 202 | 56 | 55.1 | 69.3 | 47.4 | 75.5 | 98.0 |
| 10 | 487 | 173 | 59 | 42.7 | 57.2 | 49.4 | 66.1 | 91.3 |
| 11 | 20 | 16 | 53 | 85.0 | 87.5 | 52.4 | 85.0 | 87.5 |
| 12 | 37 | 28 | 53 | 91.9 | 92.9 | 52.4 | 89.2 | 89.3 |
| 13 | 1,131 | 973 | 61 | 31.8 | 32.0 | 46.4 | 84.4 | 86.4 |
| Total | 4,628 | 2,644 | | 50.8 | 59.0 | | 77.8 | 91.6 |

Notes: *Definition 1* places the cutoff at the rounded up value of the highest (average) score observed among all treated schools in the entire region. *Definition 2* defines the cutoff value as the score that maximizes the percent of schools correctly classified across all thirteen regions (i.e., the rounded-up integer of the 95 percentile score in each region *minus 0.6 points*; see text for details). Urban, larger schools are those the Ministry of Education classifies as urban, and which have enrollments of at least 15 students in the fourth grade.

Table 3: P-900 effects on 1988-1990 and 1988-1992 math and language gain scores

| | 1988-1990 gain score | | | | | 1988-1992 gain score | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) |
| Panel A: **Mathematics** | | | | | | | | | | |
| P-900 | 2.28$^{***}$ | -0.02 | -0.11 | -0.16 | 0.25 | 3.74$^{***}$ | 1.61$^{***}$ | 1.48$^{***}$ | 1.79$^{***}$ | 2.09$^{***}$ |
| | (0.40) | (0.47) | (0.46) | (0.51) | (0.53) | (0.44) | (0.50) | (0.48) | (0.56) | (0.60) |
| Score relative to cutoff | | -0.16$^{***}$ | | | | | -0.15$^{***}$ | | | |
| | | (0.02) | | | | | (0.02) | | | |
| $\sigma_{\lambda}^{2}$ | | | 142.32$^{***}$ | | | | | 141.65$^{***}$ | | |
| | | | (18.36) | | | | | (34.01) | | |
| SES index, 1990 | | | | | 0.15$^{***}$ | | | ` | | 0.18$^{***}$ |
| | | | | | (0.01) | | | | | (0.01) |
| Change in SES, | | | | | | | | | | 0.07$^{***}$ |
| 1990-1992 | | | | | | | | | | (0.01) |
| Cubic in '88 score | N | N | N | Y | Y | N | N | Y | Y | Y |
| Region dummies | N | N | N | N | Y | N | N | N | N | Y |
| Adjusted-R$^2$ | 0.013 | 0.041 | 0.046 | 0.041 | 0.130 | 0.031 | 0.053 | 0.060 | 0.053 | 0.140 |
| Sample Size | 2,644 | 2,644 | 2,644 | 2,644 | 2,644 | 2,591 | 2,591 | 2,591 | 2,591 | 2,591 |
| Panel B: **Language** | | | | | | | | | | |
| P-900 | 4.25$^{***}$ | 0.25 | 0.18 | -0.02 | 0.54 | 5.94$^{***}$ | 2.24$^{***}$ | 2.09$^{***}$ | 1.67$^{***}$ | 2.10$^{***}$ |
| | (0.39) | (0.44) | (0.41) | (0.48) | (0.49) | (0.39) | (0.44) | (0.43) | (0.48) | (0.52) |
| Score relative to cutoff | | -0.28$^{***}$ | | | | | -0.26$^{***}$ | | | |
| | | (0.02) | | | | | (0.02) | | | |
| $\sigma_{\lambda}^{2}$ | | | 68.79$^{***}$ | | | | | 62.32$^{***}$ | | |
| | | | (5.55) | | | | | (11.21) | | |
| SES index, 1990 | | | | | 0.13$^{***}$ | | | | | 0.16$^{***}$ |
| | | | | | (0.01) | | | | | (0.01) |
| Change in SES, | | | | | | | | | | 0.07$^{***}$ |
| 1990-1992 | | | | | | | | | | (0.01) |
| Cubic in '88 score | N | N | N | Y | Y | N | N | Y | Y | Y |
| Region dummies | N | N | N | N | Y | N | N | N | N | Y |
| Adjusted-R$^2$ | 0.050 | 0.147 | 0.151 | 0.155 | 0.230 | 0.089 | 0.163 | 0.175 | 0.173 | 0.250 |
| Sample Size | 2,644 | 2,644 | 2,644 | 2,644 | 2,644 | 2,591 | 2,591 | 2,591 | 2,591 | 2,591 |

Notes: $^{***}$, $^{**}$, and $^{*}$ indicate statistical significance at the 1, 5, and 10 percent level, respectively. The sample covers urban schools with 15 or more students in the fourth grade in 1988. Huber-White standard errors are in parentheses. Columns 3 and 8 present the results from nonlinear least squares applied to the model described in the text.

Table 4: P-900 effects on 1988-1992 gain scores, within narrow bands of the selection threshold

| | ± 5 points | | ± 3 points | | ± 2 points | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Panel A: **Mathematics** | | | | | | |
| P-900 | 1.50 $^{**}$ | 1.82 $^{***}$ | 1.79 $^{**}$ | 2.00 $^{***}$ | 2.37 $^{***}$ | 2.39 $^{***}$ |
| | (0.60) | (0.66) | (0.73) | (0.77) | (0.84) | (0.85) |
| SES index, 1990 | | 0.14 $^{***}$ | | 0.13 $^{***}$ | | 0.12 $^{***}$ |
| | | (0.02) | | (0.03) | | (0.03) |
| Change in SES, | | 0.08 $^{***}$ | | 0.09 $^{***}$ | | 0.06 $^{**}$ |
| 1990-1992 | | (0.02) | | (0.02) | | (0.02) |
| | | | | | | |
| Cubic in 1988 score | N | Y | N | Y | N | Y |
| $R^2$ | 0.007 | 0.067 | 0.011 | 0.074 | 0.021 | 0.080 |
| Sample Size | 883 | 883 | 553 | 553 | 363 | 363 |
| | | | | | | |
| Panel B: **Language** | | | | | | |
| P-900 | 2.78 $^{***}$ | 2.23 $^{***}$ | 2.10 $^{***}$ | 1.96 $^{***}$ | 2.62 $^{***}$ | 2.48 $^{***}$ |
| | (0.54) | (0.57) | (0.69) | (0.70) | (0.80) | (0.75) |
| SES index, 1990 | | 0.13 $^{***}$ | | 0.12 $^{***}$ | | 0.12 $^{***}$ |
| | | (0.02) | | (0.03) | | (0.03) |
| Change in SES, | | 0.07 $^{***}$ | | 0.09 $^{***}$ | | 0.06 $^{***}$ |
| 1990-1992 | | (0.02) | | (0.02) | | (0.02) |
| | | | | | | |
| Cubic in 1988 score | N | Y | N | Y | N | Y |
| $R^2$ | 0.030 | 0.111 | 0.017 | 0.101 | 0.029 | 0.111 |
| Sample Size | 883 | 883 | 553 | 553 | 363 | 363 |

Notes: $^{***}$, $^{**}$, and $^{*}$ indicate statistical significance at the 1, 5, and 10 percent level, respectively. The sample covers urban schools with 15 or more students in the fourth grade in 1988. Huber-White standard errors are in parentheses.

Table 5: First-stage, reduced-form, and IV results for 1988-1992 gain scores, using eligibility for P-900 as an instrument

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| Panel A: First-stage estimates (**P-900**) | | | | | | |
| Eligible | 0.67 *** | 0.60 *** | 0.54 *** | 0.53 *** | 0.73 *** | 0.71 *** |
| | (0.02) | (0.02) | (0.03) | (0.04) | (0.03) | (0.04) |
| $R^2$ | 0.556 | 0.562 | 0.579 | 0.408 | 0.722 | 0.606 |
| Sample Size | 2,591 | 2,591 | 2,591 | 883 | 1,640 | 569 |
| Panel B: Reduced-form estimates (**Mathematics)** | | | | | | |
| Eligible | 3.47 *** | 1.16 ** | 1.36 ** | 1.46 ** | 1.44 ** | 1.35 * |
| | (0.38) | (0.49) | (0.58) | (0.67) | (0.69) | (0.77) |
| $R^2$ | 0.034 | 0.052 | 0.113 | 0.063 | 0.142 | 0.074 |
| Panel C: Reduced-form estimates (**Language)** | | | | | | |
| Eligible | 5.89 *** | 2.12 *** | 1.27 ** | 1.62 *** | 1.66 *** | 1.90 *** |
| | (0.34) | (0.43) | (0.51) | (0.60) | (0.61) | (0.70) |
| $R^2$ | 0.109 | 0.162 | 0.232 | 0.103 | 0.262 | 0.129 |
| Sample Size | 2,591 | 2,591 | 2,591 | 883 | 1,640 | 569 |
| Panel D: IV estimates (**Mathematics)** | | | | | | |
| P-900 | 5.19 *** | 1.93 ** | 2.51 ** | 2.75 ** | 1.96 ** | 1.91 * |
| | (0.58) | (0.81) | (1.07) | (1.23) | (0.93) | (1.09) |
| $R^2$ | 0.027 | 0.054 | 0.119 | 0.064 | 0.146 | 0.074 |
| Sample Size | 2,591 | 2,591 | 2,591 | 883 | 1,640 | 569 |
| Panel E: IV estimates (**Language)** | | | | | | |
| P-900 | 8.82 *** | 3.53 *** | 2.35 ** | 3.04 *** | 2.27 *** | 2.69 *** |
| | (0.53) | (0.72) | (0.93) | (1.11) | (0.83) | (0.99) |
| $R^2$ | 0.068 | 0.160 | 0.238 | 0.109 | 0.265 | 0.128 |
| Sample Size | 2,591 | 2,591 | 2,591 | 883 | 1,640 | 569 |
| ***Notes on all panels:*** | | | | | | |
| Linear selection term | N | Y | Y | Y | Y | Y |
| Cubic in 1988 score | N | N | Y | Y | Y | Y |
| SES controls | N | N | Y | Y | Y | Y |
| Within ± 5 points | N | N | N | Y | N | Y |
| Excluding region 13 | N | N | N | N | Y | Y |
| Sample Size | 2,591 | 2,591 | 2,591 | 883 | 1,640 | 569 |

Notes: $^{***}$, $^{**}$, and $^{*}$ indicate statistical significance at the 1, 5, and 10 percent level, respectively. The sample covers urban schools with 15 or more students in the fourth grade in 1988. Huber-White standard errors are in parentheses. The *Eligible* variable is an indicator equal to one if the school has a 1988 average test score below *cutoff definition 2* and equal to zero, otherwise. In Panels D and E, *Eligible* is used as an instrumental variable for school participation in the P-900 program.

Figure 1: Hypothetical program assignment and effects on test scores

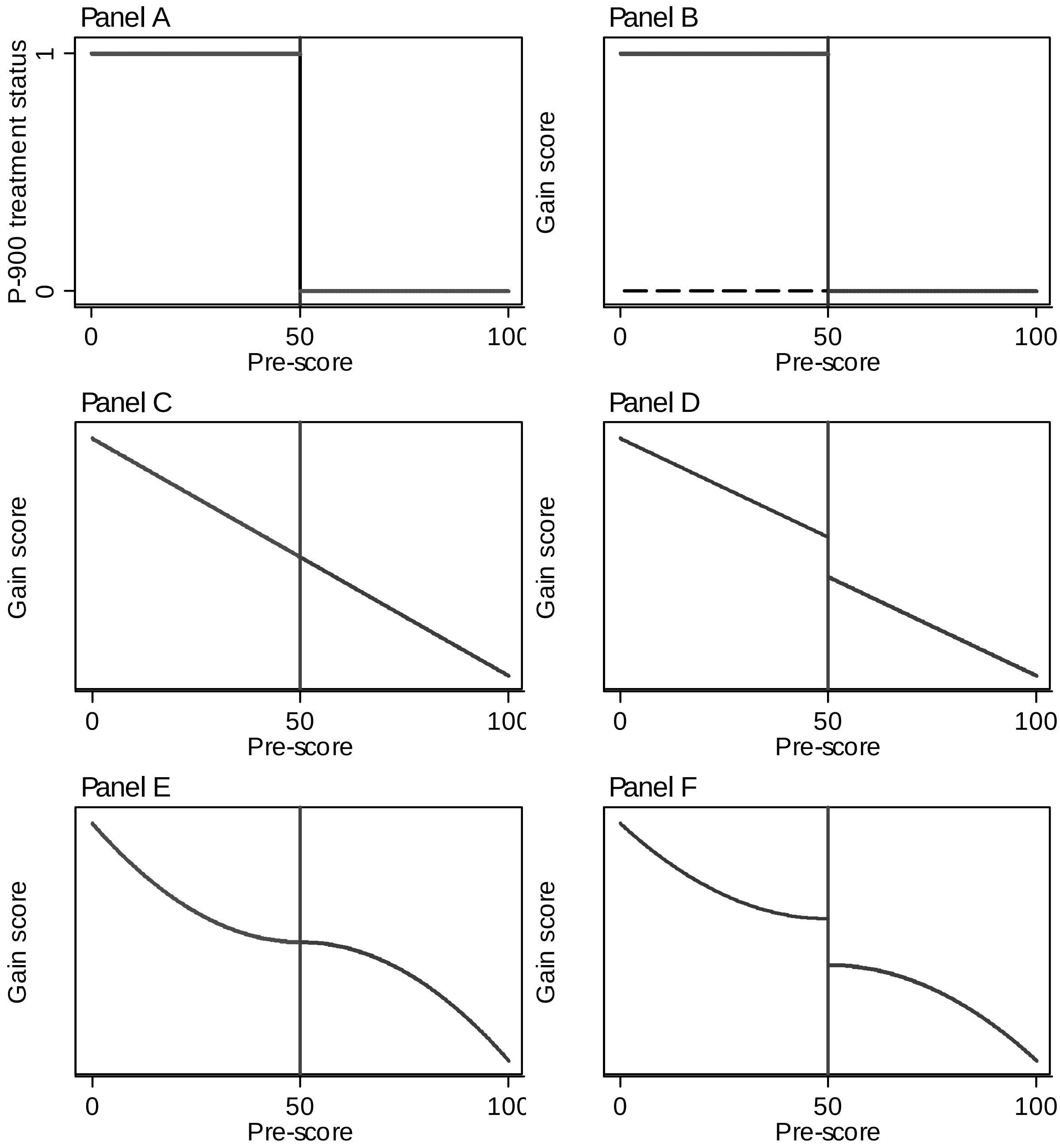
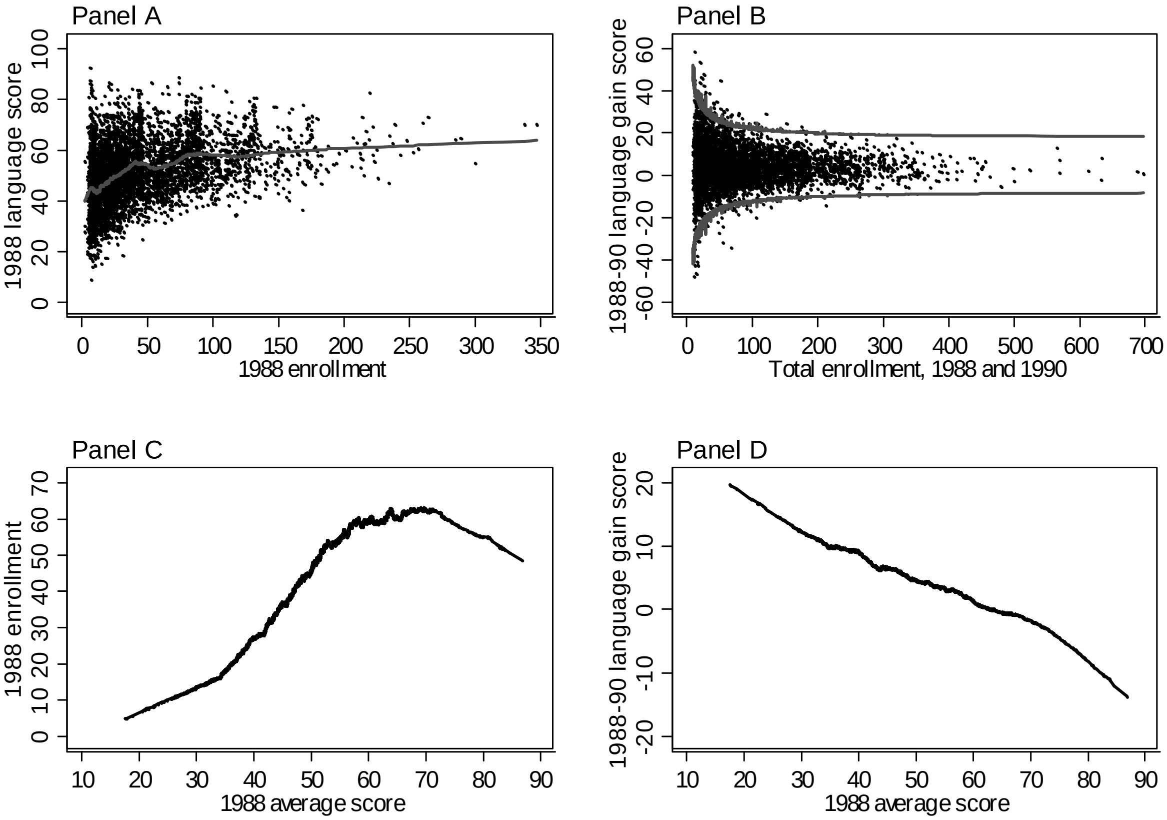
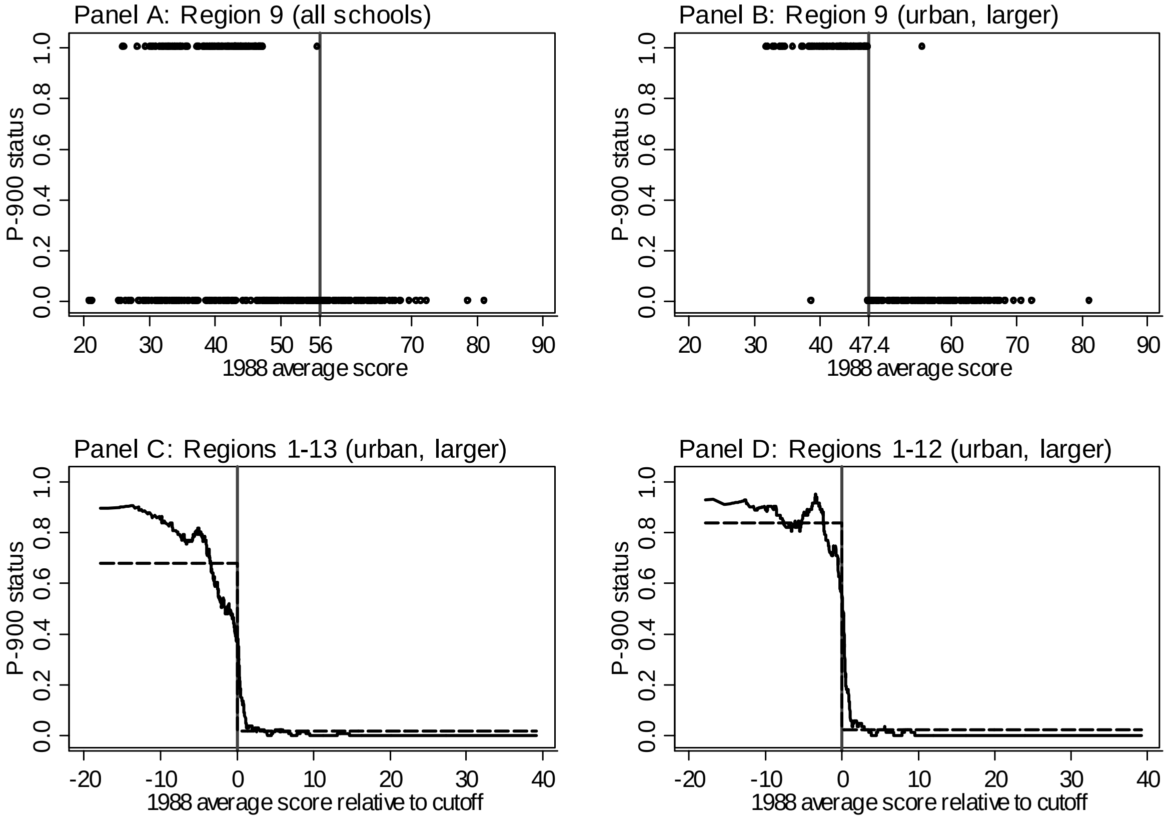


Figure 2: Average scores, gain scores, and enrollment



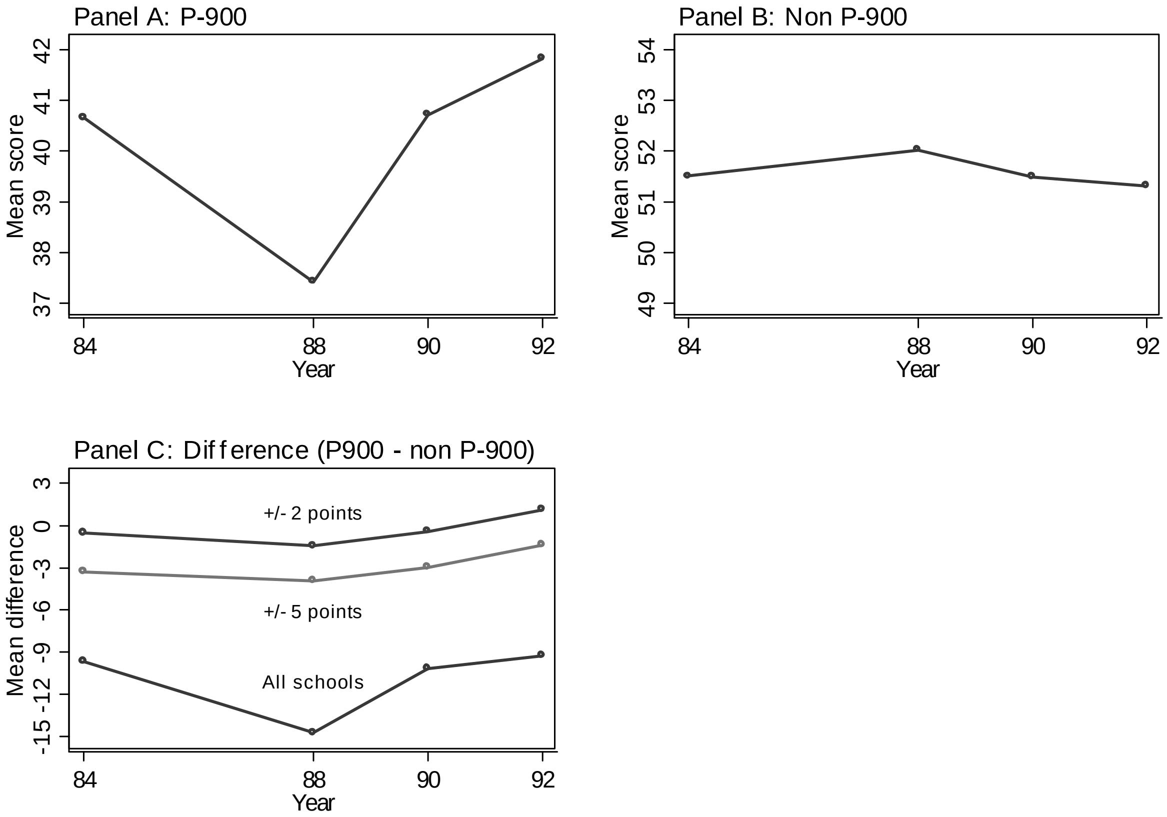
Notes: All panels use the full sample of schools. The lines in Panels A, C and D are nonparametric predictions from an unweighted local linear regression smoother with bandwidths of 0.1. The lines in Panel B are the estimated first and 99th percentiles of 1988-1990 language gain scores (see text for details).

Figure 3: Program allocation in various regions



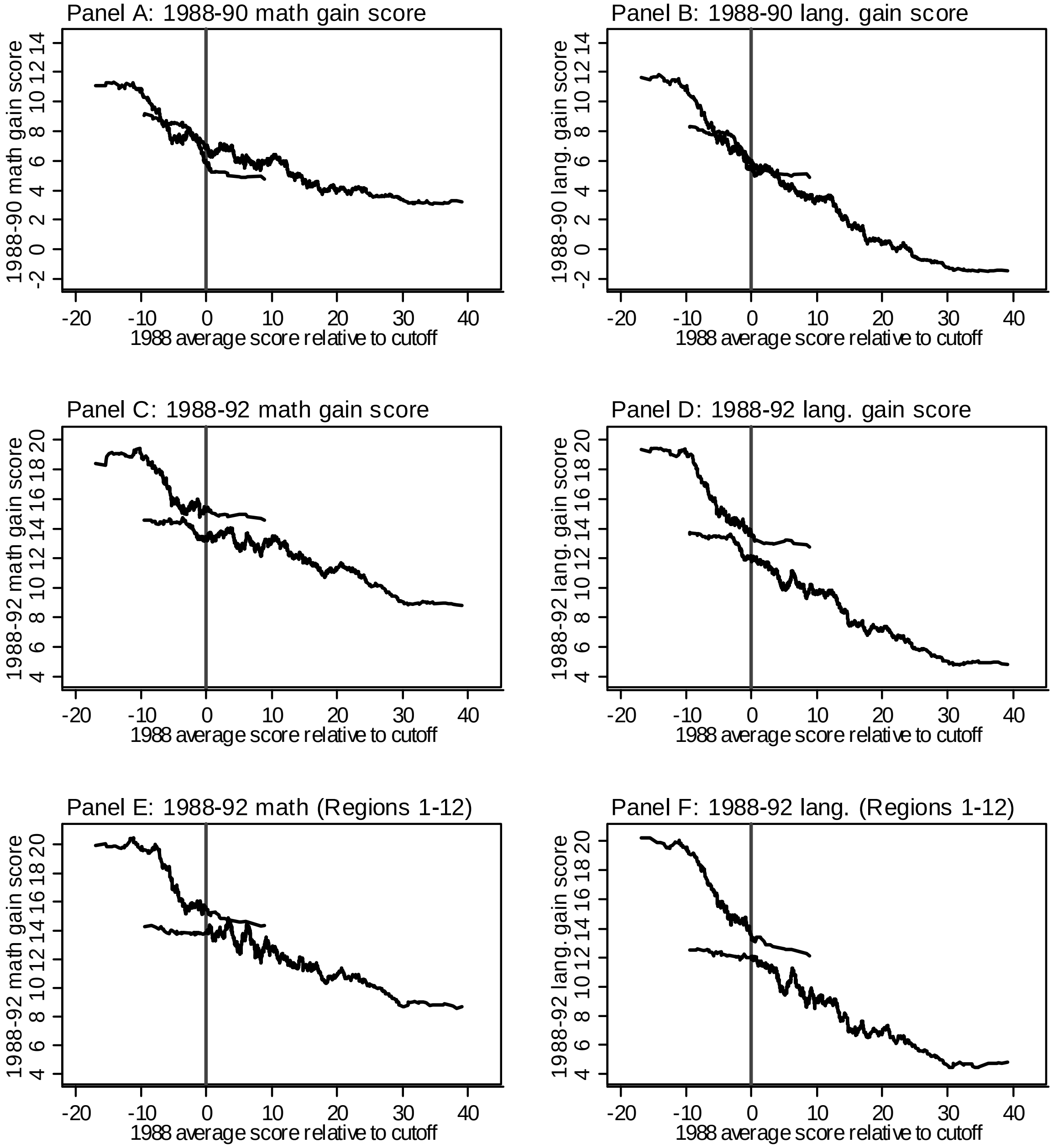
Notes: Panel A includes all schools in Region 9. Panel B includes urban, larger schools (i.e, those with fourth-grade enrollments of 15 or more) in Region 9. Panel C includes urban, larger schools in all regions (1-13), while Panel D excludes Region 13. Each dot in Panels A and B represents a school. In Panel A, the vertical line is at the rounded up value of the highest score for a school receiving P-900 in Region 9 (“cutoff definition 1” in Table 2). In Panel B, the vertical line is at the 1988 average score of 47.4 (“cut-off definition 2,” as discussed in the text and featured in Table 2). In Panels C and D, solid lines are nonparametric predictions from an unweighted uniform kernel smoother with bandwidths of 0.05. Dotted lines are OLS predictions from a regression of the P-900 indicator variable on an indicator equal to one if the school scored below cutoff definition 2 in 1988.

Figure 4: Mean language scores, 1984-1992



Notes: The figures use PER data for 1984, and SIMCE data for 1988, 1990, and 1992 (see text for details). All panels use the sample of urban schools that have at least 15 students and were tested in each year (N=1,534). The test scores in Panels A and B are standardized to a mean of 50 and standard deviation of 10. In Panel C, the top line refers to schools which had 1988 pre-scores within 2 points of their regional cut-off (N=236). The next line refers to schools which had pre-scores within 5 points of their regional cut-off (N=534). The final line refers to the full sample (N=1,534).

Figure 5: Gain scores by 1988 average score relative to the regional cutoff



Panel A: 1988-90 math gain score

Panel B: 1988-90 lang. gain score

Panel C: 1988-92 math gain score

Panel D: 1988-92 lang. gain score

Panel E: 1988-92 math (Regions 1-12)

Panel F: 1988-92 lang. (Regions 1-12)

Notes: The sample includes urban schools with fourth grade enrollments of 15 or more. Panels A-D include all regions, while Panels E-F exclude Region 13. The figures plot nonparametric predictions from an unweighted uniform kernel smoother. The bandwidths are 0.3 for the P-900 schools and 0.1 for the non-treated schools, which reflects the fact that there are over three times as many observations in the non-treated category.