The central role of noise in evaluating interventions that use test scores to rank schools

Kenneth Chay[†]
University of California at Berkeley

Patrick J. McEwan[‡]
Wellesley College

Miguel Urquiola[II]
Columbia University

October, 2003

For helpful comments we thank David Card, David Figlio, Bob LaLonde, Thomas Lemieux, Doug Staiger, and seminar participants at Berkeley, Chicago, Michigan State, Stanford, the World Bank, and the meetings of the American Education Finance Association. We are also grateful to Cristián Cox and other officials of Chile's Ministry of Education for providing data. All conclusions and errors are our sole responsibility. [†]kenchay@econ.berkeley.edu [†]pmcewan@wellesley.edu [II]urquiola@columbia.edu

Abstract

Several countries have implemented programs that use test scores to rank schools, and to reward or penalize them based on their students' average performance. Recently, Kane and Staiger (2002) have warned 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 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, these effects are much smaller than is widely believed.

1 Introduction

Several countries and U.S. states now use student test scores to explicitly rank elementary and secondary schools. A growing percentage also rely on these rankings to allocate a variety of rewards, sanctions, and assistance.[1] As a consequence, there is growing demand for information on the impact of these interventions on student performance. Recently, Kane and Staiger (2002) 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 study examines an under-appreciated implication of Kane and Staiger's findings. To the extent that 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 extremely low mean scores in a given year receive an intervention (e.g., assistance or sanctions) as a result. 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, any measured test score gains (e.g., those that would arise from a difference in differences analysis) will reflect a combination of a true program effect and spurious mean reversion. The dilemma is similar to that observed in evaluations of training programs in which assignment is based on pre-program earnings.[2]

[1] These include Israel (Lavy, 2002), Kenya (Glewwe, Ilias, & Kremer, 2003), Mexico (Santibanez, 2003), and Chile, the subject of this paper. In 2002, 43 U.S. states provided "report cards" with information on student test scores at various levels of aggregation (Kane & Staiger, 2002). Of these, 18 used the rankings to allocate rewards to schools, 20 to administer sanctions, and 28 to allocate assistance.

[2] In short, Ashenfelter's "dip" is relevant whenever treatment assignment is based on noisy pre-treatment variables, whether these are earning or test scores. (Ashenfelter, 1978; Ashenfelter & Card, 1985; Heckman, LaLonde, & Smith, 1999).

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), which since 1990 has used the mean of fourth grade test scores to identify about 900 (hence the name) of the country's lowest achieving schools. In the first three years of the program—the focus of this paper—program participation was strongly determined by whether a school's mean test score fell below given cut-off values. Participating schools received infrastructure improvements, instructional materials, teacher training, and extra 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 1988-1990 and 1988-1992 test score gains by 0.3–0.7 standard deviations; yet using P-900-type assignment rules, we can generate very similar effects during earlier periods in which the program was not yet in operation. Further, schools chosen for P-900 exhibit a sharp decline in test scores in the years preceding the program, which is consistent with a negative shock in average scores in the year used to assign program participation.

We implement a regression discontinuity approach that utilizes the discrete relation between program selection and pre-program test scores to address this problem. 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. The graphical analysis and robustness checks provide strong complementary evidence that comparing the gains of schools that fell just above and below the assignment cutoff effectively eliminates reversion biases.

We conclude that while P-900 did impact test scores, its previous favorable reviews are largely an artifact of mean reversion. Finally, the strategies illustrated in this study should be applicable to other interventions that use test-based rankings (or more generally, those that rely on some form of a "pre-score" for program assignment), especially those relying upon strict assignment rules.

2 Background on P-900

In 1990, Chile's government introduced the P-900 program, a package of interventions targeted at low-performing, publicly-funded schools.[3] The treatment encompassed four strategies (García-Huidobro, 1994, 2000; García-Huidobro & Jara, 1994). 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 1-4, 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 the in-service training and after-school workshops.

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 for schools" (Cox, 1997). On the other hand, it is possible that personnel reduced effort in the 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.

[3] About 90 percent of enrollments in Chile are in public and private schools that receive voucher-style government subsidies. All 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 further details on Chile's system of school finance, see McEwan and Carnoy (2000) and Hsieh and Urquiola (2003).

The program's assignment occurred in two stages.[4] The first relied on achievement tests administered to the population of fourth-graders in 1988.[5] 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 cut-off scores were established for each region, and schools below their region's cut-off were selected to participate.

In the second stage, regional teams of school officials reviewed each list. Some previously selected schools were removed from eligibility, based on several criteria. First, very small or inaccessible schools did not participate, in order to reduce program costs. Second, schools were excluded if they demonstrated managerial problems (such as private voucher schools that misreported their enrollments, an offense subject to financial and legal penalties). Third, there is the possibility that regional teams introduced other criteria for school eligibility.

In the past, P-900 has been lauded as a success, given the widespread perception that it substantially raised the achievement of participating schools.[6] The empirical basis of this 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 the mean increases more quickly among treated schools using a difference-in-differences framework. In a regression setting, we have

$$\Delta T_i = \alpha + \beta_1 P900_i + \epsilon_i \tag{1}$$

where ΔT_i is the change in the mean fourth-grade score of school i (hereafter the "gain score"), $P900_i$ is a dummy variable indicating treated schools, and ϵ_i is an error term. β_1 measures the gain for treated schools over and above that for untreated schools.

Table 1 reports descriptive statistics for fourth-grade language and mathematics gain

[4] For additional details, see García-Huidobro and Jara (1994) and García-Huidobro (2000).

[5] 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.

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

scores in 1988-1990 and 1988-1992. 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 this data to estimate equation (1) with 1988-1992 gain scores yields large and statistically significant estimates of β_1—equivalent to about 0.4–0.7 standard deviations depending on the test subject.[7]

3 Evaluation problems and potential solutions

For the previous 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 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 treatment effect found. We also propose a regression-discontinuity approach to address this issue.

Figure 1 (Panel A) presents a stylized version of the actual assignment rule. It plots the average pre-score of each school on the x-axis, and the treatment status, assuming a value of 0 or 1, on the y-axis. The pre-score ranges from 0 to 100, and we arbitrarily choose 30 as the cutoff. That is, all schools with pre-scores of 30 or less are treated, and the rest are not.

Panel B illustrates a simple approach to estimating P-900's effect, one that is a visual analogue to the previous estimates. The y-axis and x-axis display the gain score and the pre-score, respectively. In this case the gain score that is common to all schools is zero, and β_1 is the added gain among treated schools—the treatment effect. This interpretation is justified if pre-scores and gain scores are not otherwise related. If, on the contrary, schools with lower

[7] The precise estimates are reported later in the paper. The effects are similar, albeit somewhat smaller in magnitude, for 1988-1990 gain scores. Note that the treatment effect has a different interpretation depending on whether the outcome is the gain score for 1988-1990 or for 1988-1992. The 1990 cohort of fourth-graders in P-900 schools participated for a single year (since the 1990 test was administered towards the end of the school year). In contrast, the 1992 cohort received, at most, three years of treatment. Overall, our expectation is that treatment effects on 1988-1992 gains should be at least comparable to, and probably larger than effects on 1988-1990 gains.

scores have higher gain scores, the situation will resemble Panel C. In this extreme example, there is no identified program effect, i.e., no break in the relation between the gain and the pre-score close to the cutoff level. Nevertheless, a specification like (1) will erroneously suggest a positive treatment effect.

Why might schools with lower pre-scores have higher gain scores? The answer is rooted in the imprecise measurement of schools' mean test scores, which has two likely sources (Kane & Staiger, 2002). First, one-time events such as illness or distraction from construction noise in the school's vicinity. Second, there is sampling variation, given that each cohort of students that enters a school is analogous to a random draw from a local population. Thus, the 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 of students from which the school is drawing, and the number of students in the grade tested. We cannot assess the first of these, but we can verify the implication that scores should be more variable in schools with lower enrollments.

Figure 2 plots each school's mean 1988 score and its 1988-1990 gain score against its fourth grade enrollment. The fitted lines—unweighted smoothed values of the test scores—reveal no apparent relationship between school enrollments and mean achievement, particularly for gain scores. However, mean performance is substantially more variable among smaller schools.

This implies that some schools, especially smaller ones, probably obtained very low scores in 1988 simply because they experienced an "unlucky" circumstance (such as drawing an unusually weak group of fourth-graders in that year). They are unlikely, on average, to experience bad luck again in 1990 and 1992. Therefore, their achievement will tend to rise—i.e., they will revert towards the mean. As a result, it will be difficult to determine to what extent this improvement reflects a true treatment effect, and to what extent it is simply an artifact of mean reversion. This poses a serious challenge to any evaluation of such a program.

Fortunately, the characteristics of P-900 allow us to use a quasi-experimental regression

discontinuity (henceforth, RD) approach to address this problem.[8] Building upon equation (1), the goal is to eliminate sources of correlation, such as mean reversion, between $P900_i$ and ϵ_i. One way of doing so is to control for a smooth function of the pre-score, such as a cubic polynomial:

$$\Delta T_i = \alpha + \beta_1 P900_i + \beta_2 T_{1i} + \beta_3 T_{1i}^2 + \beta_4 T_{1i}^3 + \epsilon_i \tag{2}$$

A stricter approach is to estimate the regression within arbitrarily narrow bands close to the cut-off point. 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. Further, under this 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, Panel D depicts a stylized version of the RD approach, where the treatment effect is identified as the break in the relation between the gain and the pre-score close to the discontinuity.

Although the RD methodology provides a useful empirical lever, its application requires us to address two empirical challenges: selection and sorting. The selection issue arises because while the program's initial allocation was based on a strict assignment rule, administrators removed some of these schools from program participation—placing them in the untreated group—based on observed or unobserved attributes. This raises the possibility that assignment is correlated with unobserved determinants of achievement.

Second, sorting poses a problem because it is possible that families responded to P-900 by withdrawing their children from treated schools, or by attempting to enroll them in participating institutions, potentially altering the distribution of observed and unobserved student attributes across treated and untreated schools. A straightforward way of addressing both challenges is to include controls for schools' observable socioeconomic status (henceforth, SES) in specifications like equation (2), and examining the sensitivity of the estimates. The

[8] The RD design has recently been used to explore several issues in education (Angrist & Lavy, 1999; Guryan, 2002; van der Klaauw, 2002; Urquiola, 2000).

possibility remains of selection or sorting on unobserved variables, and below we include additional tests.

4 Program assignment

The first stage of program assignment relied on the combined mean of 1988 fourth-grade test scores, in concert with assignment cut-offs that were specific to each of Chile's 13 regions. Figure 3 illustrates the assignment rule for Region 1. In Panel A, each dot represents a school, ordered on the basis of its 1988 average score, which is on the x-axis. On the y-axis, a one indicates a P-900 school, while a zero stands for an untreated school. The vertical line is at the cut-off value corresponding to the highest test score observed among treated schools, rounded up to the nearest integer.

However, selection did not rely exclusively on 1988 test scores, which is evident in Panel A because there are several untreated schools to the left of the cutoff score. In the second stage of the assignment process, regional teams from the Ministry of Education excluded some schools even if their scores fell below the cutoff. According to administrators' accounts, this occurred mainly on the basis of school size and location.[9] Rural and small schools, in particular, were excluded in an effort to control costs. In light of this, Panel B restricts attention to urban schools with 15 or more students in the fourth grade (we will henceforth refer to these as urban, larger schools).[10] In this sample, the 1988 score induces a discrete change in the probability of treatment in Region 1, an ideal setting for an RD analysis. Of course, this judgment must be made for each region.

For this purpose, Table 2 summarizes the situation in Chile's 13 regions. Columns 1 and 2 present sample sizes, both the total and that which remains after restricting attention to urban, larger schools. Column 3 contains the cut-off score when it is set at the rounded-up integer of the highest test score observed among the P-900 schools in a region, labelled "cut-

[9] (García-Huidobro, 1994, 2000; García-Huidobro & Jara, 1994).

[10] Varying the 15 student threshold somewhat does not have a substantial effect on the results and conclusions discussed below.

off definition 1."[11] Columns 4 and 5 then present the percentage of schools that are classified correctly, i.e., a school with an average 1988 score below the cutoff did effectively receive the treatment. As expected, the cutoffs perform much better in the urban, larger school sample, where at least 90 percent of schools are classified correctly in four regions, and 80 percent or more in six regions.

Instead of disposing of regions where the discontinuity is not as stark, we rely on a second, more arbitrary definition of the cut-off. Within each region, we set it at the 95^{th} percentile of scores for P-900 schools, calling this value "cut-off definition 2". In five regions these two approaches yield the same value.[12] The results for this second definition are presented in columns 6–8. In the urban, larger school sample, at least 90 percent of schools are classified correctly in 10 regions, and 80 percent or more in all 13 regions.

We pool all this data into a single national sample. We proceed this way for expositional simplicity, and because combining the regions results in larger sample sizes. In order to combine the data, we created a variable that indicates each school's score relative to its respective regional cut-off. Figure 3 (Panel C) describes the result of this exercise for the nationwide sample of urban, larger schools. It 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 substantial changes in the probability of treatment close to the cutoff, an essential component of the RD approach. Finally, Panel D covers a subset of regions (1, 3, 4, and 6-8) in which the initial cut-off assignment correctly classifies more than 95 percent of schools. As expected, the changes in the probability of treatment are even more pronounced.

[11] This definition of the cut-off is consistent with official descriptions of the assignment process, cited earlier. According to those descriptions, schools below regional cut-offs were dropped from the treatment, but no schools above the cut-offs the were added to the treatment.

[12] These are regions with relatively few schools.

5 Results

A simple difference-in-differences analysis suggests that P-900 had a substantial effect on fourth-grade achievement. Columns 1 and 5 in Table 3 illustrate this for math and language, respectively. Panels A and B correspond to 1988-1990 and 1988-1992 gain scores. The coefficients on the treatment dummy are always statistically significant, and range between 0.3 and 0.7 standard deviations of the respective gain score distribution.

5.1 Evidence on noise and 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 was implemented. To verify this, we draw on test scores collected in 1984.[13] As a first exercise, we identified a subsample of 1,565 schools with scores available in 1984 and 1988. 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 compensatory schemes. Unless driven by mean reversion, therefore, this "fake" treatment should yield no estimated effect. In the event, estimating equation (1) with 1984-1988 math gain scores yields an estimate for β_1 of 3.5, almost exactly equal to that found for 1988-1992 (Table 3). This implies that mean reversion is indeed a first order concern in evaluating this type of program.

To provide some additional time series evidence on this issue, Figure 4 uses 1,534 schools with test scores in 1984, 1988, 1990, and 1992 (tests were not administered in 1986). Panels A and B show the annual average score of P-900 and non-P-900 schools, respectively.[14] The key observation is that scores for treated schools display a "dip" between 1984 and 1988, followed by a subsequent upward "bounce." A plausible interpretation is that many schools experienced transitory negative shocks in 1988, leading them to be selected. By 1990, mean

[13] These test scores were collected under a different system, the PER (*Programa de Evaluación de Rendimiento*), and were applied to a somewhat smaller sample of schools.

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

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 between 1984 and 1988, followed by a dip down. This is consistent with positive shocks that are followed by mean reversion. Nonetheless, the bounce and dip are less pronounced in Panel B—likely because the untreated schools are drawn from a less extreme part of the 1988 test score distribution.

In short, noise and mean reversion pose a substantial challenge to the evaluation of programs like P-900. The remainder of the paper addresses this challenge with a regression discontinuity design. It relies upon the expectation that we should observe fewer fluctuations (like those in Panels A and B of Figure 4) among schools close to regional cut-off 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 institutions (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 about 14. By 1990 however, the difference was again almost exactly equal to 10 points. Note that these differences are smaller the closer treated and untreated schools are to their regional cut-offs, and the dips are much less pronounced as well. 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, and suggests that an RD approach can be instrumental in addressing the problem of mean reversion.

5.2 Regression discontinuity results

Moving on to the RD results, Figure 5 plots schools' gain scores against their 1988 pre-scores relative to their respective regional cut-off, 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.[15] To the extent that P-900 had an effect, we should observe a break in the relationship *close to the cutoff* (one analogous to that in Figure 1, panel D). The results for 1988-1990 (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, about a fifth of a standard deviation.[16]

The regression results are consistent with the visual evidence. First, 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.[17] When columns 2 and 6 include the level of the 1988 average score as a control, the P-900 coefficients decline substantially, especially for 1988-1990 gain scores (in which case the coefficients are no longer significant). In case the regression to the mean has a more complicated form, the next two specifications include a cubic in the 1988 score, with some changes in the P-900 coefficients. Finally, columns 4 and 8 attempt to control for selection and sorting by including controls for SES, as well as regional dummies. This generally leads to slight increases in the P-900 coefficients, which are only statistically significant for 1988-1992 gain scores.[18]

We take further advantage of the RD design by limiting the sample to schools that fall

[15] In both periods the average gains are substantial, and few schools had negative gain scores. This is consistent with anecdotal evidence indicating that the test became somewhat less difficult over time.

[16] In fact two types of breaks are visible in these figures, those at the cutoff and those close by (our regression evidence below will capture both), and their magnitude is generally quite similar. This reflects the fact that the assignment discontinuities (Figure 3) are not perfect, so that there is "slippage" 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.

[17] 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.

[18] Information is available after 1992, but we make only limited use of it for two reasons. First, the program selection rules became increasingly nebulous. Some schools were removed from the treatment group and they were replaced by others. The criteria used to remove and select these schools are not well established, but it appears that selection relied more heavily on the subjective opinions of Ministry personnel and less heavily on a cut-off 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 very small scale in 1992, but rapidly expanded during the next six years to the universe of publicly-funded schools. Little is known about the rules used to allocate the program, and whether it was more or less likely to be targeted at P-900 schools. This confounds our ability to isolate the impact of P-900 in later years.

within increasingly narrow bands near the cutoff point for each region. Table 4 presents such regressions for the 1988-1992 gain scores. (We omit the similar estimates for 1988-1990 gain scores, because they simply reaffirm that P-900 had generated no effect by 1990.) There are statistically significant but modest effects in language—again, between 1 and 2 points. The point estimates for mathematics are not as consistently significant. However, it is remarkable that all point estimates continue to suggest about a 1.5 to 2 point effect, despite greatly reduced sample sizes. Finally, we return to Figure 4 to note that the graphical evidence is consistent with the regression results. Panels A and C suggested 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.

5.3 Robustness

In previous specifications, we controlled for SES to address possible biases introduced by selection and sorting. These estimates typically led to small increases in the P-900 effects, implying that selection and sorting (on observed SES) lead to small downward biases in estimates of the program effect.[19] As an additional exercise, Figure 5 (panels E and F) and Table 5 present results for schools located in regions where the cut-off correctly assigns 95 percent or more of schools, which might diminish the amount of administrator-induced selection bias. The "difference in differences" estimates (column 1) are larger than those in previous tables, which suggests that mean reversion is a particularly significant source of bias for this subset of regions. Nonetheless, the point estimates within narrow bands are quite consistent with those observed above, although they are rarely significant. This is, perhaps, because sample sizes are reduced to as few as 85 schools in this subsample.

The data also permit an entirely different identification strategy, facilitated by regional variation in cut-offs. 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

[19] In other estimates, not reported here, we assessed whether assignment to P-900 actually led to increases or decreases in the SES index of each school. The results generally suggested little change in the SES composition of each school, suggesting a small role for sorting (and, by corollary, that the previous biases are mainly produced by selection).

pre-scores *across* regions.[20] As an example, consider the sample of schools with mean 1988 scores that are greater than 50 and less than or equal to 52. In Regions 1, 3, 4, 11, and 12, these schools are below the corresponding regional cut-off (and subject to the treatment). In contrast, schools in regions 2, 5–10, and 13 are above the corresponding regional cut-off (and were 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 cut-off across regions is exogenous.[21] Table 6 focuses on 1988-1992 test scores and summarizes the results for five feasible "experiments," each conducted within a successively lower range of 1988 scores. Although the point estimates are more variable, all are positive and generally consistent with the estimates based on the RD approach.[22] Further, the stability of the estimates is consistent with the possibility that conditioning on initial test scores removes a substantial amount of the reversion bias.

To summarize, our results suggest that P-900 produced a positive effect on achievement by 1992, albeit one significantly smaller than is commonly supposed. It is also useful to mention some caveats to this conclusion. First, the absence of a program effect by 1990 could have different interpretations. It could simply indicate that the cohort of fourth-graders in 1990—only exposed to P-900 for one academic year—benefited little from the program. It could also indicate that the implemented program was qualitatively different in 1990, since some evidence suggests that fewer program elements were available in the first year. Second, we cannot identify the exact sources of this treatment effect, particularly whether it stems from the additional school resources or from the greater effort expended by "threatened" schools. It is possible that the P-900 effect represents, to some extent, gains from cheating or "teaching to the test" which cannot be assessed with available data.[23]

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

[21] It is possible that the effect of the treatment varied across regions. This impression arises from qualitative research which points to variability in the success of P-900 across regions (Carlson, 2000). For example, in several P-900 schools, principals criticized the availability of instructional materials, and some teachers claimed these arrived late or in insufficient numbers.

[22] In fact, a weighted average (weighted by the number of schools treated in every "experiment," as detailed in the notes to Table 6) of the estimates—which roughly corresponds to the average treatment effect on the treated—produces results quite close to those observed above (about 2 points in gain sores).

[23] See Glewwe et al. (2003) for a discussion of teaching to the test.

5.4 Test score noise and school rankings

So far, our results are consistent with the possibility that noisy test scores hamper the evaluation of interventions that use test-based rankings. Additionally, they might also limit their ability to identify "bad" or "good" schools. It is harder to determine the precise extent to which this happened for P-900, but we present two pieces of circumstantial evidence. First, we replicated an exercise in Kane and Staiger (2001), using 3,535 schools that report scores in four periods (1990, 1992, 1994, and 1996). We ranked schools by their mean scores in each year, and identified the lowest 20 percent (roughly akin to the actual P-900 selection rule). We then counted the number of times that each school was selected in the four years. If school performance was static and the program had no effect, then we would expect 20 percent of the schools to qualify in all four years, and 80 percent in none. If assignment were completely random, then 0.2 percent of schools would qualify in all four years, and 41 percent in none. Although we do not report the results, the proportions derived from the actual assignment more closely resemble those of a hypothetical lottery, especially for smaller schools.

Second, any difficulty in identifying under performing schools is especially relevant here because P-900 had a compensatory intent: the government desired to improve the lowest achieving schools, thereby aiding low-income children who presumably disproportionately populate such institutions. To the extent that leakage to more privileged schools did take place, however, the noisiness of test scores implies it should be greater *when they enrolled fewer students.*

Figure 6 presents evidence consistent with this. Panel A plots each school's 1988 score (relative to the regional cutoff) against its SES measure, making a distinction between treated and untreated institutions.[24] As expected there is a positive relationship between SES and a school's likelihood of scoring above the cutoff. However, many schools that score below the cutoff have high SES (as high as 100, the top of the scale). Thus, there appears to have been some "leakage" in the program to higher SES children. Panels B–D confirm that this

[24] Our data on the SES measure begins only in 1990, which is why we use that year in Figure 6.

was more severe among smaller schools, higher SES schools that were chosen for P-900 tend to be small.

6 Conclusion

In the perpetual search for policies to improve educational quality, many governments have recently turned to interventions that use test-based school rankings to allocate resources, rewards, or penalties. Not surprisingly, there is a growing demand for knowledge on the effect of these interventions. This paper has shown that noise and mean reversion induce important complications in the evaluation of such schemes. The use of intuitively-appealing evaluation schemes, like difference-in-differences, can lead to dramatically incorrect estimates of treatment effects. 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 impact, although positive, is substantially smaller than is generally believed. Most importantly, the issues and approach we use should be applicable to the evaluation of a variety of educational interventions that rely on test-based rankings.

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., & 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, O., & 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.

Carlson, B. A. (2000). *Achieving educational quality: Learning from Chile's P900 primary schools.* Serie desarrollo productivo 64, CEPAL, Santiago.

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

Gajardo, M. (1999). *Reformas educativas en America Latina: Balance de una decada.* Documentos de trabajo 15, PREAL, Santiago.

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

García-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.

García-Huidobro, J. E., & Jara, 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 Cooperacion Internacional de Chile.

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.

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.* Unpublished manuscript, Princeton University and Columbia University.

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. (2002). The promise and the pitfalls of using imprecise school accountability measures. *Journal of Economic Perspectives*, *16*(4), 91–114.

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.

Santibanez, L. (2003). *Why we should care if teachers get A's: Teacher characteristics and student achievement in Mexico.* Unpublished manuscript, RAND.

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.* Unpublished manuscript, Columbia University.

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

Winkler, D. (2000). Educating the poor in Latin America and the Caribbean: Examples of compensatory education. In F. Reimers (Ed.), *Unequal schools, unequal chances: The challenges to equal opportunity in the Americas* (pp. 113–132). Cambridge, MA: Harvard University, David Rockefeller Center for Latin American Studies.

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

Table 1: Variable means and standard deviations

Variable	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	40.7	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	0.0	1.0	0.22	0.0	1.0
Urban	0.59	0.62	0.50	0.69	0.74	0.51
Fourth-grade enrollment	44.8	46.3	39.0	50.9	54.2	39.6
	(40.5)	(42.3)	(28.5)	(40.8)	(43.1)	(28.6)
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: Test scores are expressed as the percentage of items correct. *P-900* is a dummy variable indicating program selection in 1990. *Urban* is a dummy variable indicating urban (versus rural) location. *Fourth 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 the Ministry of Education. It is scaled 0-100, with higher values indicating higher SES.

Table 2: Cut-off definitions, sample sizes, and percentage correctly classified by region

	Sample size		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	52	89.2	100.0
2	76	70	50	84.2	91.4	50	84.2	91.4
3	59	47	52	83.1	95.7	52	83.1	95.7
4	266	95	55	53.0	87.4	52	60.5	95.8
5	493	333	53	64.7	74.5	50	75.3	87.1
6	373	124	49	55.0	75.8	44	78.8	96.8
7	494	157	48	57.5	79.6	43	75.5	96.8
8	768	377	52	53.6	69.5	44	77.1	97.3
9	359	202	56	55.1	69.3	48	72.4	93.6
10	487	173	59	42.7	57.2	50	64.9	90.2
11	20	16	53	85.0	87.5	53	85.0	87.5
12	37	28	53	91.9	92.9	53	91.9	92.9
13	1,131	973	61	31.8	32.0	47	82.4	84.6
Total	4,628	2,644		50.8	59.0		76.1	90.2

Notes: *Definition 1* places the cut-off value at the rounded up highest (average) score observed among all treated schools in the entire region. *Definition 2* places the cut-off value at rounded up value of the 95^{th} percentile of the (average) score observed among all treated schools in the entire region. Urban, larger schools are those the Ministry of Education classifies as urban, and which additionally have enrollments of at least 15 students in the fourth grade.

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

	Mathematics				Language			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: **1988-1990** gain score								
P-900	2.19***	0.28	0.18	0.52	4.26***	0.67*	0.30	0.75
	(0.38)	(0.44)	(0.49)	(0.51)	(0.37)	(0.40)	(0.45)	(0.48)
	[0.30]	[0.04]	[0.02]	[0.07]	[0.61]	[0.10]	[0.04]	[0.11]
Average score, 1988		-0.14***				-0.27***		
		(0.02)				(0.01)		
SES index, 1990				0.16***				0.13***
				(0.01)				(0.01)
Cubic in 1988 score	N	N	Y	Y	N	N	Y	Y
Regional dummies	N	N	N	Y	N	N	N	Y
R^2	0.012	0.040	0.041	0.141	0.053	0.165	0.166	0.247
Sample size	2,644	2,644	2,644	2,644	2,644	2,644	2,644	2,644
Panel B: **1988-1992** gain score								
P-900	3.35***	1.59***	1.77***	1.94***	5.34***	1.96***	1.41***	1.77***
	(0.40)	(0.45)	(0.48)	(0.50)	(0.36)	(0.39)	(0.42)	(0.45)
	[0.43]	[0.20]	[0.23]	[0.25]	[0.72]	[0.27]	[0.19]	[0.24]
1988 test score		-0.13***				-0.25***		
		(0.02)				(0.01)		
SES index, 1990				0.18***				0.16***
				(0.01)				(0.01)
Change in SES, 1990-1992				0.07***				0.06***
				(0.01)				(0.01)
Cubic in 1988 score	N	N	Y	Y	N	N	Y	Y
Regional dummies	N	N	N	Y	N	N	N	Y
R^2	0.027	0.050	0.053	0.149	0.078	0.171	0.173	0.254
Sample size	2,591	2,591	2,591	2,591	2,591	2,591	2,591	2,591

Notes: ***, **, and * indicate statistical significance at the 1, 5, and 10 percent level, respectively. Regressions are weighted by the number of students in the fourth grade, and cover schools with 15 or more students at this grade level. Huber-White standard errors are in parentheses. Brackets express the P-900 coefficient as a proportion of a standard deviation of gain scores for the respective subject and time period. For 1988-90 one standard deviation in gains is 7.4 for math, and 7.0 for language. For 1988-92, these are 7.8 and 7.4, respectively.

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

	± 5 points			± 3 points			± 2 points		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A: Mathematics									
P-900	1.83***	1.45**	1.07	1.78***	1.45**	1.81**	2.05***	1.70**	1.65
	(0.54)	(0.57)	(0.72)	(0.69)	(0.72)	(0.90)	(0.74)	(0.74)	(1.08)
	[0.23]	[0.19]	[0.14]	[0.23]	[0.19]	[0.23]	[0.26]	[0.22]	[0.21]
SES index, 1990			0.17***			0.16***			0.12***
			(0.02)			(0.03)			(0.03)
Change in SES, 1990-1992			0.09***			0.09***			0.05
			(0.02)			(0.03)			(0.03)
Cubic in 1988 score	N	Y	Y	N	Y	Y	N	Y	Y
Regional dummies	N	N	Y	N	N	Y	N	N	Y
R^2	0.012	0.017	0.119	0.011	0.023	0.130	0.016	0.041	0.117
Sample size	938	938	938	560	560	560	392	392	392
Panel B: Language									
P-900	2.55***	1.35**	1.38**	1.94***	1.35**	2.02***	1.61***	1.26**	1.73**
	(0.50)	(0.49)	(0.62)	(0.65)	(0.61)	(0.77)	(0.71)	(0.64)	(0.85)
	[0.35]	[0.18]	[0.19]	[0.26]	[0.18]	[0.27]	[0.22]	[0.17]	[0.23]
SES index, 1990			0.15***			0.13***			0.14***
			(0.02)			(0.02)			(0.03)
Change in SES, 1990-1992			0.08***			0.08***			0.06**
			(0.02)			(0.02)			(0.03)
Cubic in 1988 score	N	Y	Y	N	Y	Y	N	Y	Y
Regional dummies	N	N	Y	N	N	Y	N	N	Y
R^2	0.027	0.063	0.149	0.016	0.050	0.140	0.012	0.065	0.167
Sample size	938	938	938	560	560	560	392	392	392

Notes: ***, **, and * indicate statistical significance at the 1, 5, and 10 percent level, respectively. Regressions are weighted by the number of students in the fourth grade. Sample includes urban schools with 15 or more students in the fourth grade in all regions. Huber-White standard errors are in parentheses. Brackets express the P-900 coefficient as a proportion of a standard deviation of gain scores for the respective subject. For 1988-92 one standard deviation in gains is 7.8 for math, and 7.4 for language.

Table 5: P-900 effects on 1988-1992 gain scores, in regions with at least 95% of schools correctly assigned

	All schools		± 5 points		± 3 points		± 2 points	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: **Mathematics**								
P-900	6.09***	2.41***	3.98***	1.22	2.66**	1.08	2.40*	1.44
	(0.71)	(0.91)	(0.84)	(1.41)	(1.15)	(2.18)	(1.27)	(2.47)
	[0.78]	[0.31]	[0.51]	[0.16]	[0.34]	[0.14]	[0.31]	[0.18]
Cubic in 1988 score	N	Y	N	Y	N	Y	N	Y
SES controls	N	Y	N	Y	N	Y	N	Y
Regional dummies	N	Y	N	Y	N	Y	N	Y
R^2	0.130	0.230	0.089	0.165	0.043	0.142	0.040	0.147
Sample size	490	490	215	215	117	117	85	85
Panel B: **Language**								
P-900	7.07***	2.98***	4.63***	1.81	2.89*	2.30	2.48*	2.63
	(0.67)	(0.86)	(0.84)	(1.22)	(1.18)	(1.82)	(1.30)	(2.01)
	[0.98]	[0.40]	[0.63]	[0.24]	[0.39]	[0.31]	[0.34]	[0.36]
Cubic in 1988 score	N	Y	N	Y	N	Y	N	Y
SES controls	N	Y	N	Y	N	Y	N	Y
Regional dummies	N	Y	N	Y	N	Y	N	Y
R^2	0.178	0.363	0.124	0.264	0.049	0.197	0.041	0.223
Sample size	490	490	215	215	117	117	85	85

Notes: ***, **, and * indicate statistical significance at the 1, 5, and 10 percent level, respectively. Regressions are weighted by the number of students in the fourth grade. Huber-White standard errors are in parentheses. All regressions are for the sample of urban schools with 15 or more students in the fourth grade. The sample with at least 95 percent correctly assigned includes Regions 1, 3, 4, and 6-8. Brackets express the P-900 coefficient as a proportion of a standard deviation of gain scores for the respective subject. For 1988-92 one standard deviation in gains is 7.8 for math, and 7.4 for language.

Table 6: P-900 effects on 1988-1992 gain scores, alternate identification strategy

	Mathematics			Language		
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: scores >50 and ≤52						
P-900	1.11	0.98	1.56	1.35	1.07	1.60*
	(1.20)	(1.23)	(1.13)	(0.93)	(0.94)	(0.93)
	[0.14]	[0.13]	[0.20]	[0.18]	[0.14]	[0.22]
Cubic in 1988 score	N	Y	Y	N	Y	Y
SES controls	N	N	Y	N	N	Y
R^2	0.003	0.007	0.079	0.006	0.022	0.097
Panel B: scores >48 and ≤50						
P-900	3.84***	3.87***	4.36***	2.32***	2.22**	2.68***
	(0.97)	(1.05)	(0.96)	(1.02)	(1.02)	(0.96)
	[0.49]	[0.50]	[0.56]	[0.31]	[0.30]	[0.36]
Cubic in 1988 score	N	Y	Y	N	Y	Y
SES controls	N	N	Y	N	N	Y
R^2	0.050	0.065	0.142	0.022	0.033	0.112
Panel C: scores >47 and ≤48						
P-900	3.70**	3.71**	4.41***	3.86***	3.49***	4.21***
	(1.57)	(1.56)	(1.66)	(1.14)	(1.14)	(1.20)
	[0.47]	[0.48]	[0.57]	[0.52]	[0.47]	[0.57]
Cubic in 1988 score	N	Y	Y	N	Y	Y
SES controls	N	N	Y	N	N	Y
R^2	0.045	0.075	0.112	0.060	0.084	0.131
Panel D: scores >44 and ≤47						
P-900	1.25	1.31	1.69	1.54***	1.61*	2.00**
	(1.07)	(1.10)	(1.05)	(0.92)	(0.95)	(0.88)
	[0.16]	[0.17]	[0.22]	[0.20]	[0.22]	[0.27]
Cubic in 1988 score	N	Y	Y	N	Y	Y
SES controls	N	N	Y	N	N	Y
R^2	0.007	0.011	0.071	0.013	0.016	0.087
Panel E: scores >43 and ≤44						
P-900	2.97*	2.99*	3.04*	-0.11***	0.08	0.18
	(1.71)	(1.71)	(1.62)	(1.50)	(1.38)	(1.68)
	[0.38]	[0.38]	[0.39]	[-0.01]	[0.01]	[0.02]
Cubic in 1988 score	N	Y	Y	N	Y	Y
SES controls	N	N	Y	N	N	Y
R^2	0.043	0.054	0.137	0.000	0.045	0.078

Notes: The "experiments" are as follows. For schools with test scores greater than 50 and less than or equal to 52, those in Regions 1, 3, 4, 11, and 12 are treated; Regions 2, 5–10, and 13 are untreated ($N = 221$, of which 20 are treated). For schools with test scores greater than 48 and less than or equal to 50, those in Regions 1–5 and 10–12 are treated; Regions 6–9 and 13 are untreated ($N = 193$, of which 31 are treated). For schools with test scores greater than 47 and less than or equal to 48, those in Regions 1–5 and 9–12 are treated; Regions 6–8 and 13 are untreated ($N = 116$, of which 24 are treated). For schools with test scores greater than 44 and less than or equal to 47, those in Regions 1–5 and 9–13 are treated; Regions 6–8 are untreated ($N = 237$, of which 83 are treated). For schools with test scores greater than 43 and less than or equal to 44, those in Regions 1–6 and 8–13 are treated; Region 7 is untreated ($N = 80$, of which 35 are treated). In each sample, between 9 and 44 percent of the sample consists of treated schools. We have excluded a sixth potential "experiment"—schools with test scores between 52 and 53—because it included only 4 treated schools.

Figure 1: Hypothetical program assignment and effects on test scores

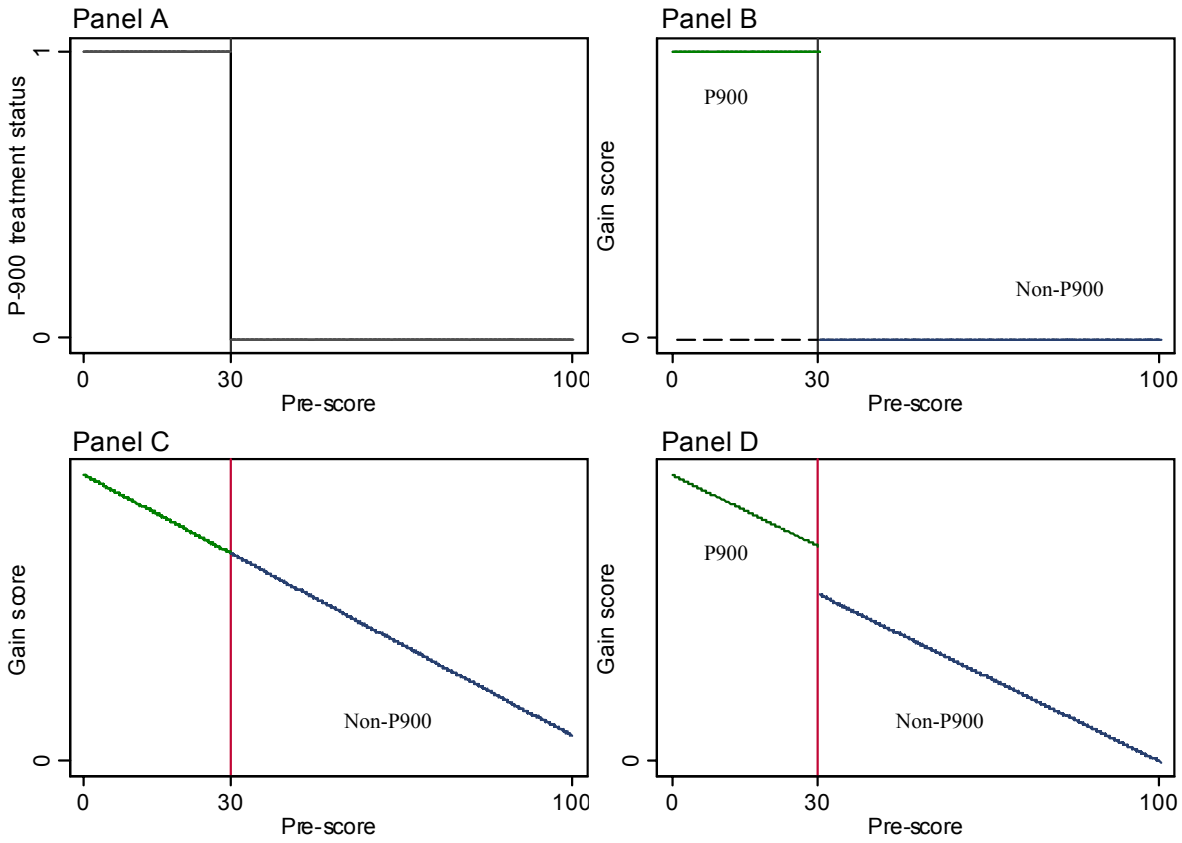
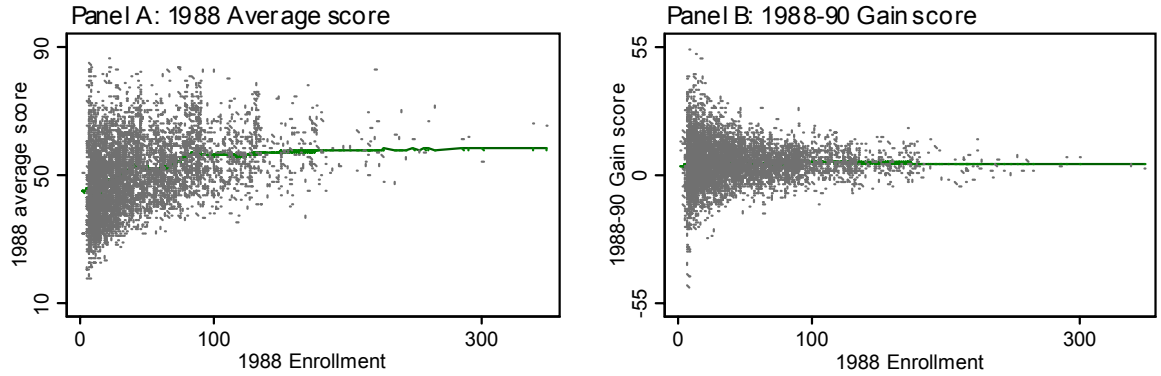
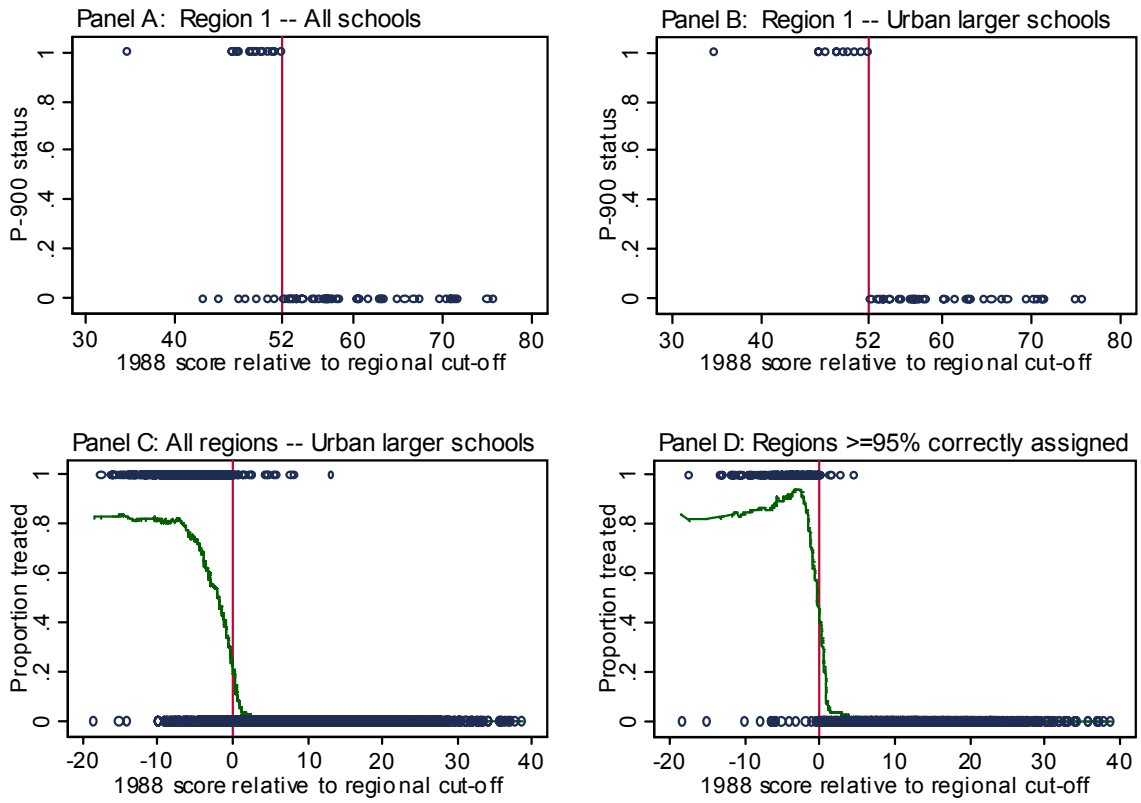


Figure 2: Average scores and gain scores by fourth grade enrollment



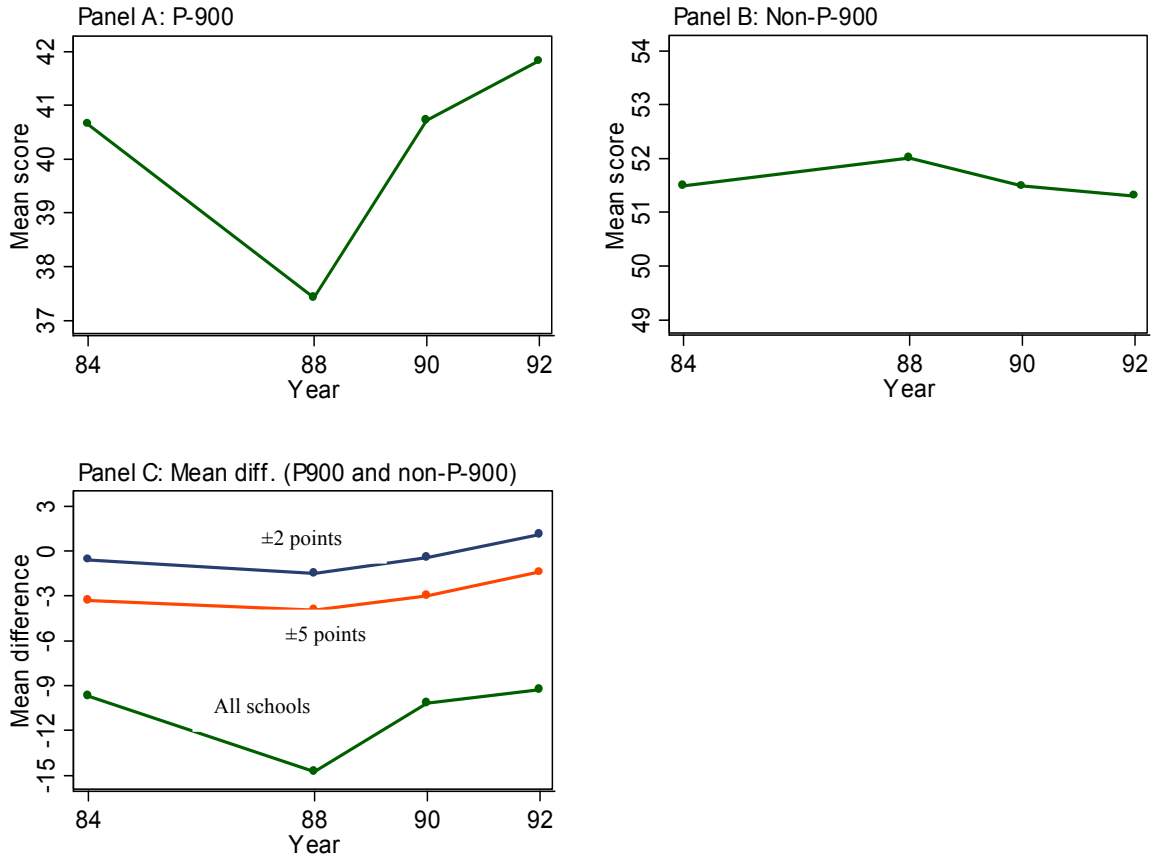
Notes: The figures cover the full sample of schools. Each dot represents a school, and the solid lines plot (unweighted) smoothed values of scores (with bandwidths of 0.10). Enrollment refers to the number of fourth-graders who took the SIMCE test in each school. The graph featuring 1988-1992 gain scores is omitted here, but it closely resembles Panel B.

Figure 3: Program allocation in Region 1 and all regions



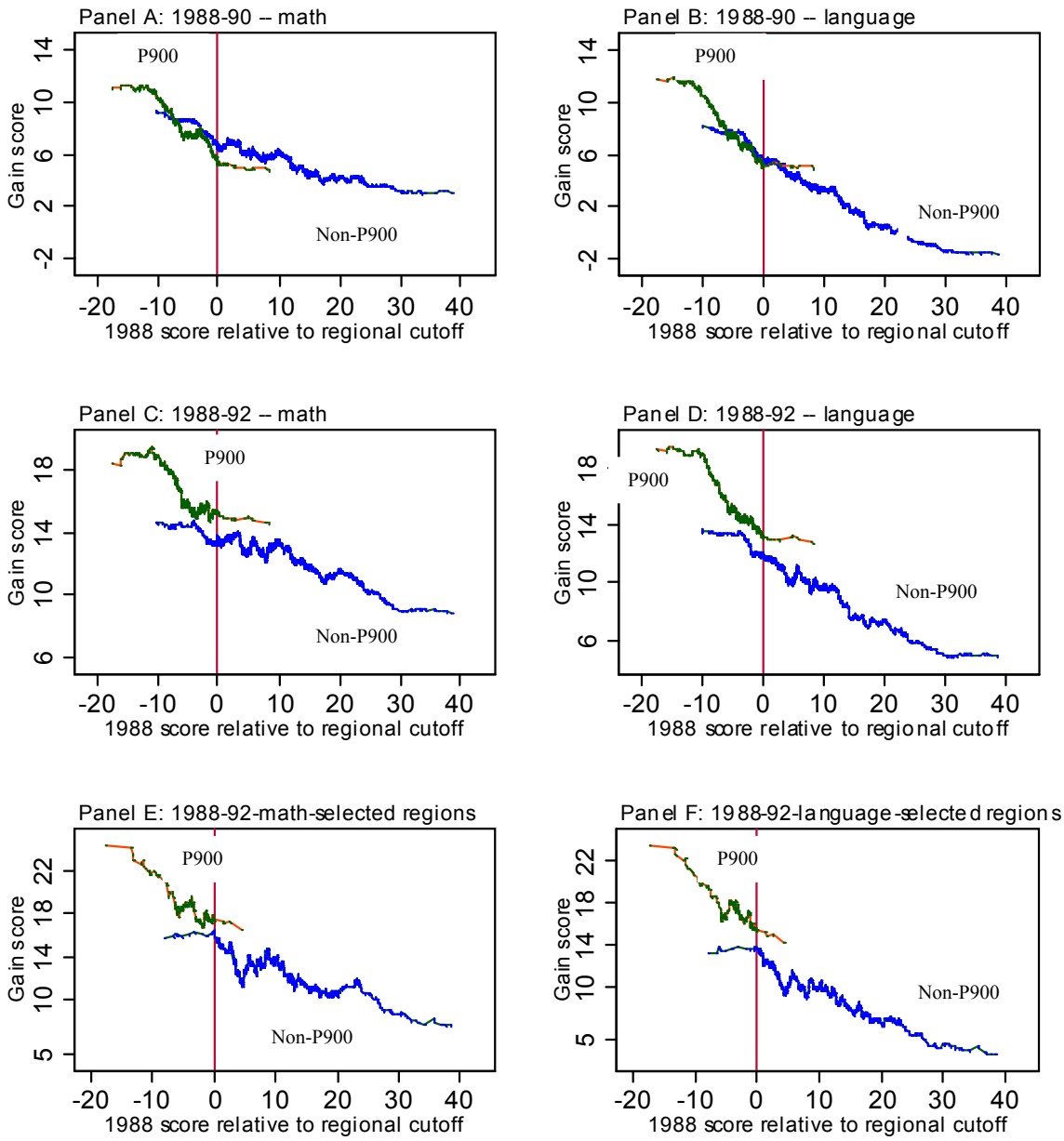
Notes: Panel A includes all schools in Region 1. Panels B, C, and D focus only on urban, larger schools – those classified as urban by the Ministry of Education, and which additionally have 4th grade enrollments of 15 or more. Panel D refers to regions in which at least 95 percent of schools were correctly assigned (as explained in the text): 1, 3, 4, and 6-8. In all panels, the 1988 score refers to the average (math and language) score for the school. Panels C and D refer to all regions, and the lines plot (unweighted) smoothed values of the proportion of schools treated (with a bandwidth of 0.10).

Figure 4: Average language scores, 1984-1992



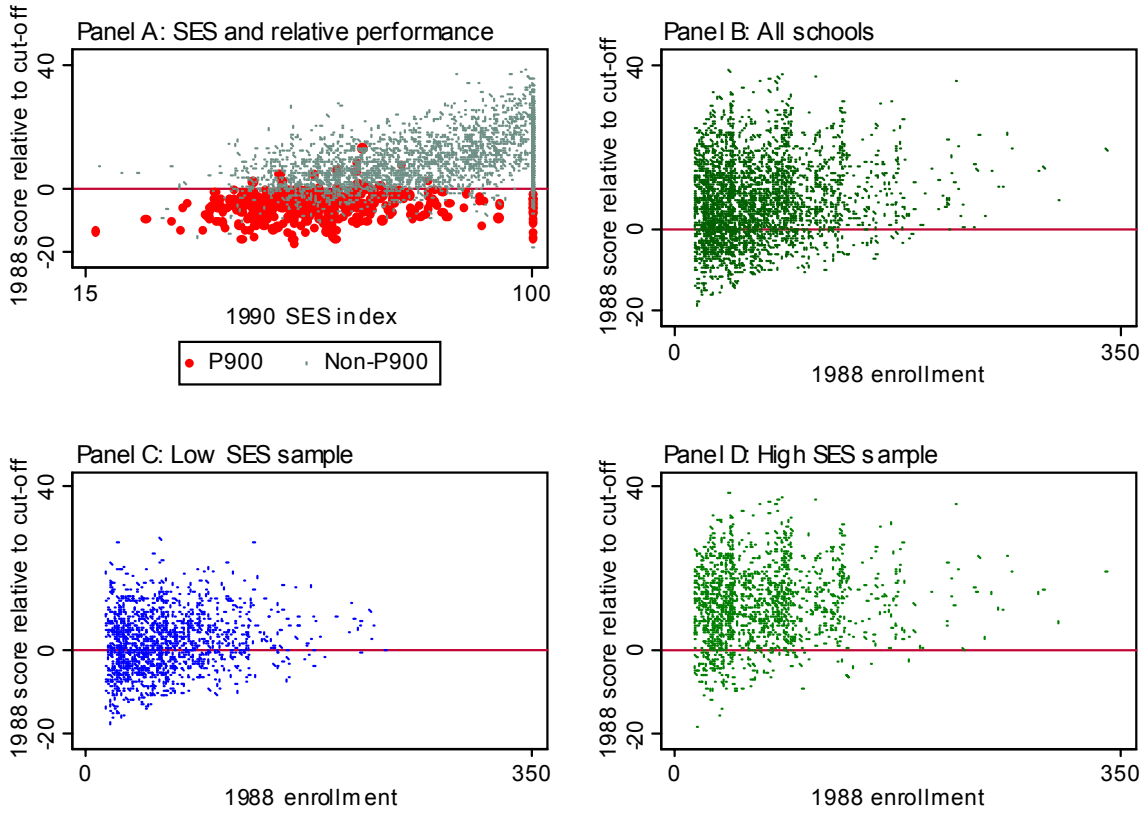
Notes: The figures use PER data for 1984, and SIMCE data for 1988, 1990, and 1992. All panels are based on a sample that includes urban schools that have at least 15 students and were in operation throughout these years (N=1,534). The test scores in Panels A and B are standardized to a mean of 50 and 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 average 1988 score relative to the regional cut-off



Notes: The figures plot (unweighted) smoothed values of gain scores. The sample includes urban schools with fourth grade enrollments of 15 or more. Panels A-D refer to all regions, and panels E and F to regions in which at least 95 percent of schools were correctly assigned—1, 3, 4, and 6-8. In all panels, the 1988 score refers to the average (math and language) score for the school. In all cases, the bandwidths are 0.3 for the P-900 schools and 0.1 for the non-treated. In part, this reflects that there are over three times as many observations in the non-treated category.

Figure 6: Noise and school rankings



Notes: The figures refer to urban schools with fourth grade enrollments of 15 or more. The SES index ranges from 0 to 100, with higher values indicating higher socioeconomic status. "Higher" and "lower" SES schools (in panels C and D) are those with values of the 1990 SES index that are above and below the median, respectively.