Tournament Incentives for Teachers: Evidence from a Scaled-Up Intervention in Chile

Author(s): Dante Contreras and Tomás Rau

Source: *Economic Development and Cultural Change*, Vol. 61, No. 1 (October 2012), pp. 219-246

Published by: The University of Chicago Press

Stable URL: https://www.jstor.org/stable/10.1086/666955

Tournament Incentives for Teachers: Evidence from a Scaled-Up Intervention in Chile

DANTE CONTRERAS
Universidad de Chile

TOMÁS RAU
Pontificia Universidad Católica de Chile

I. Introduction

Although performance-related pay for teachers has been introduced in many countries, there is still significant debate on its effects on educational performance. On one hand, advocates of teacher incentives programs argue in favor of strengthening weak incentives given that teachers are generally paid on the basis of educational attainment, training, and tenure rather than performance (Harbinson and Hanushek 1992; Hanushek, Kain, and Rivkin 1999). On the other hand, opponents argue that teachers' tasks are multidimensional and test scores do not properly reflect the performance of a given teacher. Linking compensation to test scores could cause teachers to sacrifice focusing on other skills such as curiosity and creative thinking. In addition, it may lead to different types of corruption such as gaming or teaching to the test (Neal and Schanzenbach 2010).

The body of empirical literature on the effects of teacher incentive programs based on students' performance is limited and faces some issues. First, the identification of causal effects is a complex task because of the lack of random assignment of the treatment. A teaching incentive program may be introduced because of particular characteristics of the school such as low performance, and hence, the introduction of the program may be endogenous. Second, even with

We are indebted to the useful comments of David Card, Raj Chetty, Francisco Gallego, Patrick Kline, Jeanne Lafortune, Peter Orazem, Marcela Perticará, two anonymous referees, and seminar participants at the University of California, Berkeley, Universidad de Chile, Universidad Alberto Hurtado, Pontificia Universidad Católica de Chile, Pontificia Universidad Católica del Peru, the annual meeting of the Chilean Economic Association, and the T. Paul Schultz festschrift. We also thank the funding granted by Millennium Science Initiative from the Ministry of Economy, Development, and Tourism to Micro-data Center, project NS100041. This research used as a source of information the SIMCE databases, Ministry of Education of Chile. We thank the Ministry of Education for access to that information. All the results of the study, however, are our own and are in no way committed to this institution.

a randomized treatment, the scale of these experiments is generally small, and conclusions cannot be generalized because of external validity issues.

In this article, we provide an evaluation of the only scaled-up teacher incentive program in the world. Since 1996, the Chilean Ministry of Education has incorporated a monetary-based productivity bonus called the National System of School Performance Assessment (SNED) into its standardized test scores system, the Sistema de Medición de Calidad de la Educación (SIMCE; see Mizala and Romaguera 2005). This is a rank-order tournament directed toward all municipal and private subsidized schools in the country, which represent 90% of enrolled students. This program seeks to improve teacher performance (productivity) via a monetary incentive that is allocated at the school level and awarded to teachers mainly on the basis of pupils' results on SIMCE. In the program's competitive system, schools with similar characteristics are grouped into homogeneous groups. The competition takes place within each distinct group. Thus, the SNED is a group incentive program in which schools compete against their peers on the basis of their average performance and in which monetary rewards are mainly distributed equally among all teachers in the winning schools.

Theoretically, the relationship between teacher incentives and educational performance is that the incentive may lead to increased effort on the part of the teachers, thus improving the quality of education and, hence, leading to an increase in participant schools' mean test scores. In that respect, the SNED program has some of the optimality properties described in the theoretical work of Barlevy and Neal (forthcoming), in which the authors propose an incentive pay scheme that links educators' compensations to the ranking of their students within appropriately defined comparison sets. Consistent with this suggested scheme, teachers at the same schools do not compete against each other in the SNED. The competition scheme provides incentives for effective cooperation. Additionally, teachers compete only with teachers working in similar schools (guaranteed by the homogeneous group definition).

The evidence on the effects of performance-based pay for teachers on student learning outcomes is mixed. In developing countries, evidence tends to be more favorable to these incentive programs (see, e.g., Glewwe, Ilias, and Kremer [2010] for Kenya and Muralidharan and Sundararaman [2011] for India). However, the gains tend to be short-term, which does not guarantee an increase in human capital. In developed countries, there is favorable evidence for Israel by Lavy (2002) but mixed evidence for the United States. While Figlio and Kenny (2007) find that test scores are higher in schools that offer individual financial incentives for good performance, two recent experimental evaluations report no effect of performance-based pay for teachers on student learning outcomes (Goodman and Turner [2010] and Fryer [2011] in New York and Springer et al. [2010] in Tennessee).

Contreras et al. (2005) and Gallego (2008) estimate the effect of winning the SNED on next period's test score, following a regression discontinuity strategy, finding a positive and significant effect of SNED on future test scores. However, the previous literature does not estimate the average treatment effect on the treated (ATT) of the SNED program on test scores. Learning on the effects of the introduction of the SNED on all participating schools, winning and losing, appears to be relevant especially considering that this is a nationwide policy involving 90% of schools.

We contribute to the literature on the effects of teacher incentives on academic performance by estimating the tournament effect of the introduction of the SNED, that is, the effect of the program over all schools affected by it, both winners and losers. We use an identification strategy that is basically a *matched differences in differences* between treated schools (public and private subsidized) and control schools (private fee-paying) implemented with three different empirical approaches. First, we perform a nearest-neighbor matching estimator using the methodology of Abadie and Imbens (2006) to determine the effects of the incentive on standardized test scores at the school level. Then, in the second approach, we regress the change in test scores on a set of covariates and a treatment dummy variable (affected by the tournament or not). In order to correct for potential endogeneity, we follow a double robust method, which combines inverse probability reweighting with bias adjustment incorporating the covariates included in the propensity score. Finally, we construct a panel of schools and estimate a fixed-effects model for the test scores in levels and in differences to assess the impact of the introduction of the program on test scores. Our results indicate a significant tournament effect on participant schools of between 0.14 and 0.25 standard deviations for language and math test scores.

The rest of this article is organized as follows. Section II provides a brief description of the SNED teaching incentive program. The methodology and empirical strategy are discussed in Section III. Section IV describes the data. The results are presented in Section V. In Section VI, we present our conclusions.

II. The Program

Chile has had a decentralized school system since the reforms of the 1980s when the administration of public-sector schools was transferred from the Ministry of Education to municipalities. This reform opened the way for the private sector to participate as a provider of publicly financed education by establishing a voucher-type, per-student subsidy. Thus, in Chile, schools are divided into three school administration types, based on funding source and administration: (*a*) public schools with public funding and administration, (*b*) private state-subsidized schools in which the financing is provided by the state but with private administration, and (*c*) private fee-paying schools in which both funding and administration are provided by the private sector. The voucher system gives families complete freedom to choose schools for their children. They can choose a subsidized school, either municipal or private. Alternatively, they can choose a fee-paying private school.[1]

SNED is directed at all primary and/or secondary subsidized schools in the country and is financed by the government. Note that private fee-paying schools are excluded. In the year 2000, 90% of all schools in Chile were municipal or publicly subsidized private schools. The SNED, which is a supply-side incentive, was created with two objectives. First, it was intended to improve educational quality provided by state-subsidized schools through monetary rewards to teachers. This strategy, defined as a pay-for-productivity wage compensation, sought to change the fixed salary structure. The second objective was to provide information about school progress to the school community, parents, and those responsible for children. It was expected that the school administrations and teachers would thus receive feedback on their administrative decisions and teaching.[2]

The SNED program is defined as follows. Schools are grouped by region. Then they are classified according to location (urban/rural area) and as primary or secondary schools. Once these groups are defined, they are then subcategorized by vulnerability and socioeconomic characteristics according to the official classification provided by the Ministry of Education: high, medium-high, medium, medium-low, and low levels. The ministry refers to the sets of associated schools as homogeneous groups and investigates differences inside each group. This method is used because it is considered inappropriate to compare the performance of schools with adverse external conditions, such as low parental educational level, low family income, and high social vulnerability, with the performance of schools with good external conditions. Therefore, following a tournament design, the competition among schools takes place within each homogeneous group.[3]

Once the group has been defined, the SNED index is computed for each school within its homogeneous group and the schools are ranked according to

[1] The school choice is limited by the school selection criterion and tuition fees. For a discussion, see Contreras, Sepulveda, and Bustos (2010).

[2] See Mizala and Urquiola (2007) for an evaluation of the effects that being identified as a SNED winner has on schools' enrollment, tuition levels, and socioeconomic composition.

[3] According to Mizala and Romaguera (2005), the classification within a homogeneous group has remained relatively stable, except for some changes in the methodology between rounds 1 and 2.

this index. Top schools, accounting for 25% of the enrollment in each homogeneous group, are chosen for the Teaching Excellence Subsidy. These funds are distributed directly to the teachers as follows: 90% of the total bonus goes directly to all teachers on the basis of the number of hours worked. Schools allocate the other 10% as extra bonuses for those teachers whose contributions were noteworthy. Payments are made quarterly. For the 1996–97 SNED competition, the yearly amount per teacher at awarded schools was about US$370. This is approximately 40% of a teacher's monthly income, equivalent to an annual salary increase of 3.33%.[4]

The factors determining the SNED index are the following:

1. Effectiveness, which is the educational results achieved by the school in relation to the population served: This considers the average SIMCE score in both language and mathematics during the most recent evaluation. For the 1996–97 SNED competition, this variable corresponded to the 1995 SIMCE score in eighth grade and the 1994 SIMCE score in fourth grade. This factor was weighted to 40% in that year's SNED index but has now been decreased to 37%.
2. Improvement, which consists of the differential in educational achievement obtained over time by the school: It was weighted 30% in the 1996–97 SNED and then decreased to 28% in the following rounds. This measure of improvement varies on the basis of the previous SIMCE score at the school level. For schools whose previous SIMCE test was in fourth grade, this variable measures the average difference between the 1992 and the 1994 SIMCE scores. For those schools whose previous test was in eighth grade, the comparison considered was between 1993 and 1995.
3. Initiative, defined as the capacity of the school to incorporate educational innovations and involve external agents in its teaching activities: It is measured through educational projects, teaching workshops, agreements with institutions and/or companies for work placement, and other related activities. The source used for this indicator is the SNED survey. It has a weight of 6% in all SNED rounds.
4. Improvement of working conditions and operations of the school: The indicators that make up this factor are the permanent teaching staff and substitute teachers. This factor is weighted to only 2% for all SNED rounds.
5. Equality of opportunities, which consists of school access by pupils, class retention, and the inclusion of pupils with learning difficulties: It is mea-

[4] The monetary incentive has increased to about US$1,000 per year in the 2006–7 round, which is about 80% of a teacher's monthly salary.

TABLE 1
DESCRIPTION

Factor	SNED Weighting 1996–97 (%)	SNED Weighting 1998–99 (%)
Effectivity	40	37
Improvement	30	28
Initiative	6	6
Improvement of working conditions	2	2
Equality of opportunities	12	22
Incorporation of parents	10	5

Source. Ministry of Education.

sured through retention rates, the inclusion of multideficit and severe deficit students, integration into development projects, and the pass rate of students. The information is obtained from the enrollment and performance statistics of the Ministry of Education and from the SNED survey. The weight for this index was 12% in the 1996–97 round and increased to 22% afterward.[5]

6. Integration and participation of teachers and parents in the development of the school programs and initiatives: This factor is calculated from two indicators. The first is the establishment of parental centers and the second is the acceptance of their work. This information comes from a questionnaire for parents of the SIMCE students and the SNED. This factor had a 10% weight in the 1996–97 round and then decreased to 5% in the following rounds.

Each of these factors consists of a series of indicators. The indicator with the greatest relative weight is the SIMCE scores, representing 70% of the 1996–97 SNED index. Table 1 shows the evolution of those proportions.

III. Evaluation and Identification Strategy

In order to evaluate the effect of SNED on test scores, we address the following question: How does competition for the prize increase, if at all, schools' mean test scores? According to the neoclassical models of incentives, the introduction of a tournament may change the incentive structure of teachers, and competition for the prize may be reflected in increased effort and, hence, an increase in participant schools' mean test scores.

This question is not trivial given the difficulties faced when trying to identify a causal relationship. The construction of a valid control group given the design of the program is troublesome. Participating schools in the SNED tournament

[5] This component prevents the possibility of selecting only good students.

account for 90% of schools in Chile (private fee-paying schools being noneligible). It is likely that pretreatment characteristics for private fee-paying schools in the control group would be different from those of subsidized schools in the treatment group. We take three approaches to address this issue. The first approach we pursue is to construct a control group using a matching procedure and perform a matched difference-in-difference approach. This implies the choice of an algorithm to match treated and control observations such as nearest neighbor or matching in the propensity score. For nearest-neighbor methods, such as those used by Abadie and Imbens (2006), it is not clear how to choose the number of neighbors. In addition, in the case of using propensity score methods, a misspecified propensity score may lead to bias in the treatment effects estimates. A second alternative is to pursue a double robust method. These methods have the advantage of being robust to either a misspecified propensity score or model. A third approach we follow in this section is a panel data estimator for the ATT. This alternative allows us to exploit the panel structure of our sample and to control for time and fixed effects.

In order to examine heterogeneous response to the treatment, we study if the tournament implies the presence of schools that are always "on the money" (top schools that systematically rank in the upper quartile or so) and schools that are "out of the money." If this is the case, only a reduced number of schools in the treated group would actually be affected by the tournament. We propose a simple method to identify schools on the money by estimating the probability of winning the 1996–97 tournament with pretournament data and then computing the difference between actual and predicted test scores for groups with different probabilities of winning.

A. Matching in Characteristics

The first approach we follow is a matched differences in differences. This approach basically combines a matching algorithm with differences in differences and thus addresses any unobserved characteristics that are constant across time between the two groups. The algorithm we use is the one proposed by Abadie and Imbens (2006), which is a nearest-neighbor approach and the matching by characteristics. The characteristics chosen in the implementation of the matching procedure are average parent education, region dummies, and student-teacher ratio (STR).

B. Double Robust Methods

The second approach to shed light on the tournament effect on test scores is to implement a *double robust* estimator. This method was first introduced by Robins and Rotnitzky (1995) and consists of estimating a weighted regression of the outcome variable on the treatment dummy and the covariates. The weights are computed as a function of the propensity score. The advantage of the method type is that the estimator is consistent whenever one of two things happens: the model is correctly specified or the propensity score is correctly specified.

The propensity score will be calculated by estimating a probit regression for the probability of being treated, that is, the probability of being a private voucher or public school against being a private school. With the results of the probit, we can obtain the propensity score and the weights. As described in Busso, DiNardo, and McCrary (2009), double robust methods allow us to estimate average treatment effect (ATE) and ATT by adjusting the weighting scheme. We will focus on ATT effects given that the program affects 90% of schools.

The weighting scheme we use to estimate the ATT is given by $w = (p_s/(1 - p_s))/(\hat{p}/(1 - \hat{p}))$ for untreated and $w = 1$ for the treated, where p_s is the propensity score and $\hat{p}$ is the unconditional probability of being treated. This is the scheme IPW1 analyzed by Busso et al. (2009). Then, we estimate the following weighted regression model:

$$\Delta Y_{i,t} = X_{i,t-1}\beta + \alpha d_{i,t} + \epsilon_{i,t}, \tag{1}$$

where $\Delta Y_{i,t}$ is the difference between SIMCE test scores before and after the introduction of the program, $X_{i,t-1}$ are the covariates related to characteristics of the school before the treatment, $d_{i,t}$ corresponds to a dummy variable that is equal to one if the school participates in the program (public and private subsidized) and is equal to zero if the school is private, and $\epsilon_{i,t}$ is the residual.

The variables in the propensity score should reflect the differences in characteristics between private fee-paying schools and subsidized schools (public and private). As we can see in table 2, public and private subsidized schools are very similar in terms of average parental schooling but are very different from private schools for this measure. We control, then, for average parental education. We also include regional dummies to better capture the heterogeneity of localization of these schools. Then we add the STR since it has been reported that public and private subsidized schools tend to have larger class sizes than private fee-paying schools, which is highly correlated with the STR (Urquiola and Verhoogen 2009). Finally, we add a dummy variable equal to one if the school is a full-day school to control for differences in school day length between treated and untreated schools.

In order to avoid comparability issues between different test scores, we standardize each measure subtracting the mean and dividing by the standard deviation of the control population. This implies that we will be able to identify

TABLE 2
SCHOOL CHARACTERISTICS, ADMINISTRATIVE DEPENDENCY, AND PERFORMANCE

	1996			2006		
Variables by School	Private	Private Subsidized	Public	Private	Private Subsidized	Public
SIMCE score:						
Mathematics	83.61	69.59	65.61	288.09	243.75	231.47
	(7.31)	(11.42)	(10.21)	(28.34)	(34.10)	(31.94)
Language	84.48	70.38	65.43	289.33	252.84	243.77
	(6.14)	(11.74)	(9.78)	(24.84)	(28.45)	(28.44)
Household variables:						
Average schooling of parents	4.44	2.70	2.18	4.01	3.20	2.69
	(.58)	(.76)	(.47)	(.20)	(.69)	(.52)
Average schooling of mothers	. . .	. . .	. . .	4.01	3.20	2.72
				(.18)	(.70)	(.52)
Average schooling of fathers	. . .	. . .	. . .	4.14	3.23	2.71
				(.37)	(.68)	(.54)
Average household income (thousand CLP)	. . .	. . .	. . .	1,045.55	290.86	148.10
				(205.95)	(196.71)	(82.13)
School variables:						
Rural	.01	.16	.51	.04	.21	.60
	(.12)	(.36)	(.50)	(.19)	(.41)	(.49)
Average number of students taking the test	43.15	56.94	46.18	35.28	38.61	25.60
	(33.17)	(49.59)	(44.26)	(30.83)	(36.86)	(30.06)

Source. Authors' calculation based on SIMCE data set. CLP = Chilean pesos.
Note. Standard errors are in parentheses.

only how the treated group does relative to the control population, but note that this is in the spirit of the design.

This first approach will not exploit the panel data nature of the data and will measure only the effects of the program after the first round. A second alternative is to construct panel data and is explored below.

C. Panel Data Estimation

In addition to performing the matching and double robust estimation, we construct a panel of schools from 1990 to 1999. This allows us to measure the tournament effect after two rounds of the SNED while controlling for pretreatment information, school fixed effects, and geographic region trends.

The construction of the panel is not straightforward since the available data are school averages until 1997 and are at the individual level afterward. Hence we have to compute school averages and create some aggregate variables such as average parental schooling, type of school, and so forth. Given the continuous changes in the questionnaires, it is a challenge to create or keep track of certain variables such as STR or parental schooling categories. However, we are able to create an unbalance panel with more than 6,500 schools with an average number of periods of 6.4 (minimum of 1 and maximum of 10).

The equation we estimated is

$$Y_{i,t} = \alpha_1 treated_{i,t} + \alpha_2 after_{i,t} + \alpha_3 treated_{i,t} \times after_{i,t} + X_{i,t}\beta + \gamma_1 trend_{i,t} + \sum_{j=1}^{12} \gamma_{2j} region_{i,j} \times after_{i,t} \times trend_{i,t} + \epsilon_{i,t}, \tag{2}$$

where $treated_{i,t}$ is a dummy taking the value one if the school is eligible for SNED (public or private subsidized) and zero if it is private fee paying. The dummy $after_{i,t}$ is a binary variable equal to one if it is 1996 or later. The dummies $region_{i,j}$ are 12 geographic region dummies. The variable *trend* is a time trend and $\epsilon_{i,t} = \eta_i + u_{i,t}$. The parameter of interest is the one accompanying the interacted dummies for eligibility after the SNED started. As the error structure shows, we estimate this equation with school fixed effects and school clustered standard errors as well. Hence, we are able to identify the parameters for time-variant covariates. In addition to the school fixed effects, we have dummies per level of parental schooling, and we add geographic region trends interacted with the before-after dummy. We perform this estimation in levels and first differences, unweighted and then reweighted in the same way as in the previous section.

One concern is the effect of other programs that may be confounded with those of the SNED. The only program that we think might confound with SNED is the Jornada Escolar Completa (JEC), which was a program to increase the length of the school day from a half day to a full day. This program was launched in 1997, a year after the first round of SNED, and included 19% of schools. We control for this by including a dummy variable for the length of the school day.

D. Heterogeneous Response to the Treatment

In order to examine heterogeneous response to the program, we estimate the probability of winning by simulating the tournament with pretreatment information. Then we compare the deviation of actual and predicted test scores with pretreatment information across the predicted probabilities of winning. Even though we do not have a full mapping from the probability of winning to the optimal effort that a teacher should exert, we interpret this as a reflection of the variation in the "power" of the incentives.

We follow a procedure similar to that used by Neal and Schanzenbach (2010). We compute the mean difference between test scores predicted by pre-SNED data with actual test scores following the introduction of the program. We perform this strategy separating by groups according to their probability of winning the tournament.

In order to determine this probability, we simulate the SNED tournament with pretreatment information. Then we estimate a linear model of the 1996 index on the lagged value of math test scores and its second difference as follows:

$$sned_{i,t} = \beta_1 simce_{i,t-1} + \beta_2 \Delta_{i,t-1} simce + \beta_3 \Delta_{i,t-2} simce + \beta_4 X_{i,t} + \epsilon_{i,t}. \qquad (3)$$

These variables capture the level and improvement factors defined in the formula of the SNED index. Given that we do not have the rest of the data tracked by the SNED index, such as equality of opportunity, we add additional controls such geographic region and urban/rural dummies.[6] Then we predict the SNED index and compute the probability of winning for each school in its homogeneous competition group. This is done by computing the cumulative distribution after sorting the schools (ascending) by the predicted SNED index in each homogeneous group. Thus, our measure of the probability of winning is the percentile in which each school is located in the cumulative distribution of the predicted SNED index in its respective homogeneous group.[7]

To discover the presence of schools on the money and out of the money, we compare the post-tournament test scores with our predictions. The distribution of this "prediction error" across the probability of winning (computed with pretreatment data) may indicate the presence of schools on the money and a tournament effect for at least a subpopulation of eligible schools.

In order to do this, we construct a panel data set of eligible schools for 1989–95. Then we estimate a linear dynamic panel data model of test scores on characteristics (such as school size, parental schooling, expenditure in tuition, and lags of the dependent and independent variables) following Arellano and Bond (1991).

With our estimated model, we predict the 1996 test scores and compute their deviation from the actual 1996 test scores. Hence, we can observe the distribution of this prediction error across the previously computed probability of winning. The presence of sure losers would be reflected in the presence of a

[6] These estimates may suffer from omitted variable bias given the previous discussion. However, we are interested in the rankings that they generate since we simulate the tournament and winners are determined by the top 25th percentile. Given that the predicted score is linear in the parameters, the bias induced by the omitted variables would act as a monotonic transformation of the "true" prediction and would not change the ranking in a particular homogeneous group.

[7] It is important to remark that schools do not move on the basis of performance. According to Mizala and Romaguera (2005), homogeneous groups have remained relatively stable.

marked (fat) lower tail. Conversely, the presence of sure winners would be reflected in the presence of an upper tail. Since this particular prediction error is between the post-tournament 1996 test scores and the results predicted with pretournament data, if the tournament was ineffective, the prediction error and the probability of winning would not be related.

It is important to mention that this exercise could reflect other dimensions of heterogeneity as well that are different from the power of the incentive. Quality of the managerial team and principal or even persistent positive shocks at the school level could affect the results. However, it still is an interesting exercise to perform to assess the presence of heterogeneity, especially in a scaled-up incentive program. This may be informative about which types of schools are driving the results.

IV. Data

This article uses data from the national SIMCE (1989–99) test. Tests are conducted for students in fourth and eighth or tenth grade depending on the year. We have aggregate data at the school level for 1989–97. However, since 1998, student-level data are available. We use school-level data since the tournament is at the school level. SIMCE data sets also include information on family and school characteristics. The continuous changes in the questionnaires during the period analyzed limit the availability of covariates for the estimation. However, we are able to construct geographic region dummies, dummies for parent education level, and STR.[8]

Table 2 presents the main school characteristics and performance levels by administrative school type: public, private subsidized, and private fee-paying. The table summarizes information for the years 1996 and 2006. It indicates that private fee-paying schools have students of higher socioeconomic status than private subsidized and public schools. Private fee-paying schools show the highest average household income and parental education levels. School performance in mathematics and language is consistently lower in public schools compared to private subsidized and private fee-paying schools.

It is important to note that there was a change in the SIMCE scoring scale in 1998. In 1996, the SIMCE test had an average score around 70 points with a standard deviation of about 10 points. Then the SIMCE test switched to a scale with an average of 250 points with a standard deviation of 50 points. Since 1998, SIMCE tests have been comparable over time, using the same scale and grading.

[8] For STR we are able to construct the series for years 1993, 1994, 1995, and 1996.

TABLE 3
SCHOOLS PERFORMANCE: WINNERS AND LOSERS

	1996		2006	
Variables by School	Winners	Losers	Winners	Losers
SIMCE score:				
Mathematics	68.27	66.24	249.37	248.27
	(11.19)	(10.52)	(28.44)	(25.73)
Spanish	68.49	66.33	257.11	255.92
	(11.28)	(10.37)	(24.12)	(22.84)
Household variables:				
Average schooling of parents	2.38	2.33	3.04	3.15
	(.67)	(.59)	(.58)	(.61)
Average schooling of mothers	. . .	. . .	3.03	3.16
			(.59)	(.61)
Average schooling of fathers	. . .	. . .	3.05	3.15
			(.59)	(.60)
Average household income (thousand CLP)	. . .	. . .	232.26	250.25
			(159.81)	(159.66)
School variables:				
Rural	.43	.40	.36	.30
	(.49)	(.49)	(.48)	(.46)
Average number of students taking the test	48.50	49.85	41.83	45.43
	(46.14)	(46.06)	(36.11)	(37.47)

Source. Authors' calculation based on SIMCE data set. CLP = Chilean pesos.
Note. Standard errors are in parentheses.

Table 3 summarizes the variables discussed above for winning and losing schools. This information is presented for 1996 and 2006. In both years we do not observe any significant differences in educational performance and socioeconomic characteristics between winning and losing schools. At first sight, the results look random, but they should be interpreted carefully. First, given that competition occurs within a homogeneous group, we expect to observe similar socioeconomic characteristics among schools in a particular group. Second, the simple average in performance is not capturing differences between homogeneous groups. In other words, given that competition occurs within groups, differences in performance need to be observed among schools in the same homogeneous group.

Table 4 shows the distribution of schools according to the number of awards received over time. We restrict the sample to schools that compete in all six rounds of SNED so that the frequencies are not influenced by different compositions of schools in different rounds. This table shows that 38% of schools have never been awarded the SNED bonus. Only a small fraction of schools have won the SNED several times. In other words, according to the evidence, there might be some schools that are out of the money or that have teachers who do not respond to the program with higher effort (heterogeneity in teacher type). One explanation for this is that it may be due to measurement error and/

TABLE 4
SCHOOLS BY NUMBER OF AWARDS (PARTICIPANTS IN SIX ROUNDS)

Number of Awards	Frequency	Percent
0	3,108	38.64
1	2,085	25.92
2	1,339	16.65
3	802	9.97
4	427	5.31
5	215	2.67
6	68	.85
Total	8,044	100

or mean reversion, which may be important in determining the winners. It is important to mention that this is not a problem for the tournament effect. In fact, in a symmetric tournament, one would expect a higher effort but the winner being arbitrary.

V. Results

In this section we present the results of the evaluation strategies discussed in Section III for the evaluation of the tournament effect using three different empirical approaches: nearest-neighbor matching, double robust methods, and fixed-effects panel data estimation.

The set of covariates considered are STR, region dummies, full-day dummy, and average parental education dummies (primary, secondary, and college). The treatment group includes public and private subsidized schools and the control group is the private schools.[9]

In table 5 we can see the results for the ATT for math and language scores using Abadie and Imbens (2006) matching. The outcome variable is the 1995–96 pairwise difference in standardized test scores. We also use the 1995–97 pairwise difference in standardized test scores since schools might have taken some time to react to the introduction of the program. The effect is positive, ranging from 0.11 to 0.19 standard deviations in language and from 0.19 to 0.44 standard deviations in math. When focusing on the 1995–96 difference, the results are not significant in some specifications when the number of neighbors is low, such as for language. On the other hand, results for math test scores are robust to the number of neighbors and to the difference considered.

[9] We also considered excluding the public schools from the treatment group in order to increase comparability between the treatment and control groups, obtaining similar results and a slightly better balance of covariates after reweighting. However, by excluding public schools we would be losing schools that are important from a policy perspective.

TABLE 5
TOURNAMENT EFFECTS IN MATH AND LANGUAGE, ABADIE-IMBENS MATCHING

	Language	Math
Four neighbors:		
Difference 1995–96	.11	.23
	(.12)	(.15)
Difference 1995–97	.19*	.29***
	(.11)	(.12)
Six neighbors:		
Difference 1995–96	.16	.31***
	(.11)	(.12)
Difference 1995–97	.18***	.43***
	(.09)	(.08)
Eight neighbors:		
Difference 1995–96	.15*	.19*
	(.09)	(.10)
Difference 1995–97	.17***	.44***
	(.07)	(.07)
Observations	4,138	4,190

Note. Standard errors are in parentheses.
* Significant at 10%.
** Significant at 5%.
*** Significant at 1%.

As mentioned in Section III, an alternative approach to evaluating the tournament effect on test scores is to estimate equation (1) by using double robust methods, which consist of reweighting by a function of the propensity score and adding the covariates included in the propensity score estimation. These methods have been reported to perform well in finite samples and have the advantage of being robust to a misspecification of the propensity score or the model separately.

In order to implement a double robust method, we need to construct a weight based on an estimate of the propensity score as discussed in Section III. The estimation of the propensity score can be found in the appendix (tables A1–A4) as well as the weighted means for treated and untreated. It was obtained by running a probit of the treatment status on average parent education, geographic region dummies, STR, and a full-day school dummy. It appears to be very well specified (pseudo $R^2 = 0.73$, sensitivity above 97%, and specificity above 83%). The covariates are relatively well balanced since the pairwise mean difference tests for each covariate reject the null hypothesis of equality in only five of 16 cases. However, the joint hypothesis of equality is rejected.

When implementing the double robust method, we used Eicker-Huber-White robust standard errors to account for heteroskedasticity, and in the panel data estimates, we used school fixed effects and school clustered standard errors as well.

TABLE 6
TOURNAMENT EFFECTS IN MATH AND LANGUAGE, DOUBLE ROBUST METHOD

	ATT	SD	*t*-Test	*N*
Language:				
Difference 1995–96	.15	.16	.93	1,807
Difference 1995–97	.26***	.11	2.51	1,786
Math:				
Difference 1995–96	.23	.23	1.02	1,816
Difference 1995–97	.24***	.08	2.95	1,786

Note. Eicker-Huber-White robust standard errors.
* Significant at 10%.
** Significant at 5%.
*** Significant at 1%.

Table 6 presents the ATT for math and language test scores. Similarly to the nearest-neighbor matching, the outcome variables are the 1995–96 and 1995–97 pairwise differences in standardized test scores. The effect is positive but insignificant when using the 1995–96 pairwise difference as the outcome variable. When the outcome variable is the 1995–96 pairwise difference in standardized test scores, the effects are 0.15 and 0.23 standard deviations for language and math test scores, respectively, but are not significant statistically. When using the 1995–97 difference, the effects are stronger and statistically significant at about 0.24 and 0.26 standard deviations for math and language tests, respectively.

The panel data evidence is consistent with what was found with double robust methods, as can be seen in table 7. The variable *treated* is the treatment dummy and *after* is a dummy variable that takes the value of one from 1996 on. Since the estimation is school fixed effects, only time-variant covariates are identified (regional dummies are excluded from the estimation). In column 1, we present the results for math scores controlling by school fixed effects and region dummies interacted with the before and after dummies and a trend. Thus, we allow for different trends before and after the introduction of the SNED. In column 2, we perform the same fixed-effects estimation for math test scores but reweighted the same as in the double robust approach. In columns 3 and 4, we do the same as in columns 1 and 2 but for language scores. As we can observe, for math and language scores, we find a significant effect of about 0.16 and 0.14 standard deviations, respectively, when no weights are introduced. The effect rises to 0.25 and 0.22 standard deviations when we reweight. To explore the presence of heterogeneity between public and private subsidized schools, we add a dummy variable for public schools interacted with the before-after dummy (public × after). The results show no significance for the interacted dummy. Separated estimations were also performed, that is,

TABLE 7
TOURNAMENT EFFECTS (LEVELS), PANEL 1990–99

	Math		Language	
Coefficient	(1)	(2)	(3)	(4)
After	.177***	.067	.210***	.147***
	(.036)	(.058)	(.033)	(.055)
Treat × after	.157***	.252***	.138***	.220***
	(.022)	(.033)	(.020)	(.039)
Public × after	.001	−.002	−.007	−.008
	(.014)	(.015)	(.013)	(.014)
Primary	.027	.032	−.001	.010
	(.049)	(.050)	(.040)	(.041)
Secondary	.066	.058	.074*	.095**
	(.050)	(.052)	(.042)	(.047)
College	.062	.055	.085**	.070
	(.049)	(.051)	(.041)	(.044)
Full-day	.072***	.072**	.056***	.065**
	(.016)	(.032)	(.014)	(.030)
Constant	−.094*	.127***	−.090**	.108***
	(.048)	(.049)	(.040)	(.040)
School fixed effects	Yes	Yes	Yes	Yes
Region trends	Yes	Yes	Yes	Yes
Reweighted	No	Yes	No	Yes
Observations	43,270	43,270	43,231	43,231

Note. Standard errors are in parentheses. Standard errors are clustered by school. School fixed effects are included. The outcome variable is in levels.
* Significant at 10%.
** Significant at 5%.
*** Significant at 1%.

public against private and private subsidized against private, finding no significant difference.

In table 7 we also see a dummy variable controlling for full-day schools, which is positive and significant, ranging from 0.06 to 0.07 standard deviations for language tests, depending on the specification, and 0.07 standard deviations for math. The results are very similar to those found by Bellei (2009), especially for language scores. For math tests, his results vary from 0.00 to 0.12 standard deviations, “with 0.07 standard deviations being the most convincing estimate” (637). Finally, the variables measuring average parental schooling show positive but nonsignificant results.

Now, we perform the panel data estimation with the outcome variable in first differences. The interpretation of the results here is different since, when estimating in differences, we are testing if the “trajectories” are different after the program (pointing to an acceleration of gains over time). When estimating in levels, we are instead testing if the levels of SIMCE scores are higher in treated schools after the start of the SNED program. These results can be observed in table 8. The effects of the SNED on trajectories are positive and significant for

TABLE 8
TOURNAMENT EFFECTS (FIRST DIFFERENCES), PANEL 1990–99

	Math		Language	
Coefficient	(1)	(2)	(3)	(4)
After	−.282***	−.729***	−.056	−.295
	(.078)	(.144)	(.073)	(.194)
Treat × after	.153***	.286***	.047*	.046
	(.027)	(.064)	(.026)	(.055)
Public × after	.016	−.002	−.010	.029
	(.014)	(.027)	(.014)	(.021)
Primary	.029	.035	−.015	−.006
	(.106)	(.108)	(.083)	(.085)
Secondary	.019	−.065	.010	.014
	(.109)	(.141)	(.086)	(.092)
College	.020	−.118	−.043	−.275
	(.115)	(.215)	(.092)	(.197)
Full-day	.070**	.149**	.007	.059
	(.029)	(.072)	(.026)	(.090)
Constant	−.048	.025	−.02	.028
	(.106)	(.125)	(.082)	(.096)
School fixed effects	Yes	Yes	Yes	Yes
Region trends	Yes	Yes	Yes	Yes
Reweighted	No	Yes	No	Yes
Observations	29,315	29,315	29,271	29,271

Note. Standard errors are in parentheses. Standard errors are clustered by school. School fixed effects are included. The outcome variable is in differences.
* Significant at 10%.
** Significant at 5%.
*** Significant at 1%.

math testing, ranging from 0.15 to 0.29 standard deviations when we reweight. For the language test, the effect on trajectories is smaller, about 0.06, and non-significant at the 1% level.

As explained before, the SNED program has two main objectives: incentives for effort and the provision of feedback to teachers and administrators. These two objectives might be confounded in the interpretation of our results in the panel data model, since our results incorporate more than one round of the SNED. Muralidharan and Sundararaman (2010) test whether the diagnostic feedback has an impact on student learning with experimental evidence from India. They found that there was no impact of the feedback and monitoring on student learning outcomes when comparing feedback (treated) and no feedback (control) schools. Given that in their performance pay study (Muralidharan and Sundararaman 2011) the treated group gets feedback, the combination of their two papers suggests that the impact comes from the incentives and not from feedback. We use these results to suggest that our estimated effects are most likely due to the incentives.

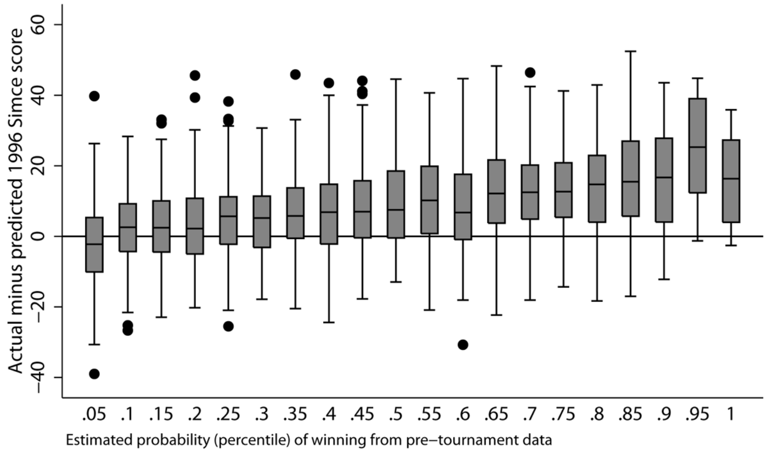

Figure 1. Box plots of the test score prediction errors across probability of winning groups: all eligible schools.

Heterogeneous response to the program. Now we analyze if the program provided the same magnitude of incentives for schools with different ex ante probabilities of winning. In figure 1, we can see box plots of the prediction error of test scores across the percentiles of the predicted probability of winning.[10] It can be seen that the tournament seems to affect schools with a probability of winning greater than the 60th percentile. This suggests the existence of a large fraction of schools that do not respond to the incentive program. On the other hand, we observe a positive and significant tournament effect for schools in the 65th percentile and higher. The last group, schools in the 95th percentile, still shows a positive effect lower than previous percentiles and with lower statistical power.[11] This lack of monotonicity is expected since optimal effort is not a monotonically increasing function of the probability of winning (e.g., why exert more effort if you are certain to win?). For an example of this, see Goodman and Turner (2010) and Fryer (2011) on the New York City teacher bonus program.

Now, in order to see if noneligible schools show the same pattern, we repeat the exercise for private schools (false experiment). Then we predict their SIMCE test score for 2006 using pretreatment information and compute the probability of winning on "artificial" homogeneous groups. These groups

[10] In case the reader is not familiar with this type of plot, each box contains 50% of the data for each category, from the 25th to the 75th percentile. The line in the middle of the box represents the median or 50th percentile, and the other lines (whiskers) are 1.5 times the interquartile ratio (distance from the 25th to the 75th percentile). Observations lying outside the whiskers are considered outliers.

[11] We compute the *t*-statistics for schools in the 95th percentile and above, finding a *p*-value equal to .23.

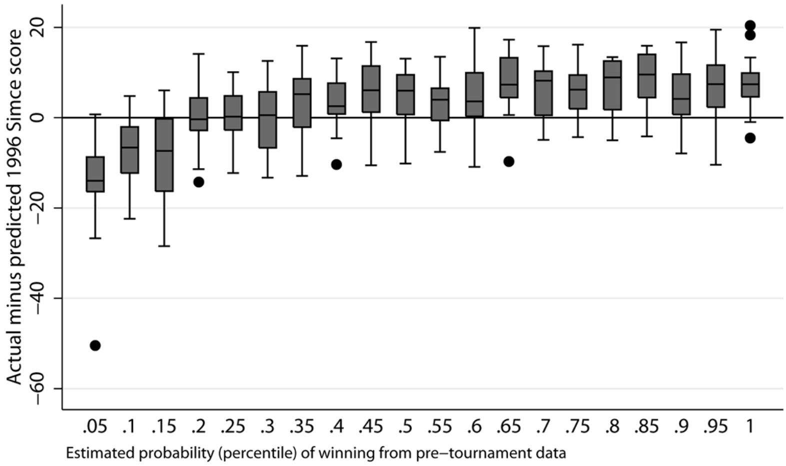

Figure 2. Box plots of the test score prediction errors across probability of winning groups: noneligible schools.

were constructed using geographic region and urban/rural status; the empirical probability of winning is computed for each group. Figure 2 shows the box plots of the prediction error of test scores across the predicted probability of winning. It is interesting to note that the pattern observed in figure 1 is not observed here. Thus, noneligible schools are not subject to the tournament. This validates our identification strategy.

Gallego (2008) estimates the effects of the predicted probability of winning the next round of SNED on current test scores, finding a significant effect of 0.08 standard deviations, with larger effects on schools with low socioeconomic status and low initial results. His results imply monotonicity of test scores on the probability of winning. However, optimal effort is not a monotonically increasing function of the probability of winning as mentioned above. Our results are different: we find an effect only in a subset of schools, which responds to the incentive in comparison to other schools.

Finally, we estimate the double robust model for schools on the money according to our observations in figure 1. Hence, we restrict the estimates to schools with a probability of winning between the 60th and 95th percentiles. The results are presented in table 9. We see that the ATT rises for all estimates, increasing the statistical significance. Of course this is a reduced number of schools; thus these results are not generalizable to the whole population but indicate that schools on the money may feel higher tournament effects than those out of the money.

Related to the previous findings, even though this is a scaled-up program, our control group is a subset of the schools (private fee-paying); hence there

TABLE 9
TOURNAMENT EFFECTS IN MATH AND LANGUAGE, DOUBLE ROBUST METHOD, EXCLUDING SURE LOSERS AND SURE WINNERS

	ATT	SD	*t*-Test	*N*
Language:				
Difference 1995–96	.32*	.18	1.77	621
Difference 1995–97	.37***	.15	2.45	628
Math:				
Difference 1995–96	.42*	.21	1.94	621
Difference 1995–97	.32***	.11	2.98	628

Note. Excludes schools with probability of winning less than .60 and more than .95.
* Significant at 10%.
** Significant at 5%.
*** Significant at 1%.

might be external validity issues to consider. Specifically, if private subsidized schools or public schools that are similar to private fee-paying schools are driving the results, the policy implications cannot be generalized to public schools. If this subset happens to be highly correlated with on-the-money schools, this enhances the heterogeneous effect results across the predicted probability of winning and thus is less generalizable to overall public policy.

We have one final thought about the cost-effectiveness of this program. Cabezas, Cuesta, and Gallego (2011) analyze a sample of different types of programs worldwide such as teaching incentives, scholarships, class size reductions, textbooks, full-day schools, and so forth. They find that the cheapest programs in their analysis cost as low as $2 per 0.1 standard deviation increase in language test scores, including the individual incentive program in India and the SNED in Chile. Moreover, the full-day school program (JEC) implemented in Chile costs about $636 per 0.1 standard deviation increase in language test scores. Thus, teaching incentive programs are very cost-effective in increasing language test scores in the short run compared with other programs.

VI. Conclusions

This article contributes to the literature on incentive pay for teachers based on school academic performance. We examine the effect of a rank-order tournament, the National System of School Performance Assessment, on standardized test scores. A major feature of this program is that it is scaled up and affects about 90% of Chile's schools. This is particularly important since the evidence reported in the previous literature has been based on pilot experiences with only a small number of affected schools. We provide evidence for the only scaled-up teaching incentive program in the world.

We compare public and private voucher schools to private fee-paying schools following a matched difference-in-difference approach with three empirical implementations. Matching, double robust methods, and panel data estimation are pursued in order to estimate the tournament effect of the introduction of the program on test scores. We find a significant effect of the program on standardized math and language test scores. The results are comparable to those found by Muralidharan and Sundararaman (2011) in India with the distinction that the SNED includes 90% of schools, whereas the experiments in India affected only a small fraction of schools. The results are robust to different approaches and vary between 0.16 and 0.25 standard deviations for math and from 0.14 to 0.26 for language scores in double robust and panel data specifications.

One common concern is that schools may focus on attempting to maximize their SNED score instead of focusing on student progress. This has been mentioned by Carnoy et al. (2007), who argue that given that the SIMCE test is an intercohort test, schools might have an incentive to simply increase performance in the tested grades. They show that when intracohort measures are compared, most of the awarded schools show little, if any, progress in terms of academic achievement. However, as pointed out by Gallego (2008), Carnoy et al.'s estimates do not control for mean reversion, and when mean reversion is controlled for, there is a positive and significant intracohort gain.

A second concern is that the increase in test scores may not represent a real increase in human capital because of gaming by the teachers toward testing instead of overall learning (Koretz 2002; Glewwe et al. 2010). Probably the only real way to address this concern is to have a different measure of learning outcomes outside of the tested measures (such as in Muralidharan and Sundararaman [2010, 2011]). Unfortunately, we do not have such a measure, but it would be an interesting contribution to include other measures of learning besides test scores in the following rounds of the SNED program to check for short-term gaming.

The empirical evidence presented in this article also provides support for educational policies oriented toward greater differentiation in the salary structure for teachers. In many countries where teachers unions are very important (in particular in Latin America and less developed countries), a wage structure that recognizes pay for productivity would be theoretically efficient. This article provides evidence supporting such a wage structure as a mechanism to increase student achievement. However, this article also shows that these types of tournaments are productive for only a specific subset of schools. Thus, the evidence shows that such a rewards system may create improvements in only a fraction of schools. These findings are particularly relevant since they are based on results from the only scaled-up program in the world. Therefore, further research on SNED is needed to evaluate different designs and incentive mechanisms aimed at affecting a broader range of schools.

Appendix
Supplementary Tables

TABLE A1
PROBIT, PROPENSITY SCORE. DEPENDENT VARIABLE: TREATED

Variable	Coefficient
Primary	−4.003***
	(.707)
Secondary	−6.232***
	(.710)
College	−8.294***
	(.741)
Region_2	.796*
	(.468)
Region_3	1.660***
	(.472)
Region_4	.763**
	(.389)
Region_5	−.004
	(.291)
Region_6	.135
	(.363)
Region_7	.525
	(.378)
Region_8	.403
	(.319)
Region_9	.938**
	(.415)
Region_10	.222
	(.398)
Region_11	−.757
	(1.057)
Region_12	.403
	(.455)
Region_13	−.015
	(.280)
Student-teacher ratio	.010**
	(.004)
Full-day	−.794***
	(.273)
Constant	5.882***
	(.760)
Observations	4,195
Pseudo R^2	.73

Note. Specificity 84%, sensitivity 97%.
* Significant at 10%.
** Significant at 5%.
*** Significant at 1%.

TABLE A2

BALANCE OF TREATED AND UNTREATED USING REWEIGHTING

	Untreated		Treated		
	Mean	SD	Mean	SD	*t*-Statistic*
Primary	.83	.38	.81	.39	.33
Secondary	.14	.35	.16	.37	.35
College	.03	.17	.01	.08	2.01
Region_2	.24	.43	.04	.19	6.19
Region_3	.00	.03	.02	.14	.89
Region_4	.01	.08	.04	.19	1.00
Region_5	.05	.22	.15	.36	1.79
Region_6	.05	.23	.04	.20	.32
Region_7	.01	.11	.04	.20	1.02
Region_8	.16	.37	.11	.32	.94
Region_9	.01	.09	.05	.21	1.14
Region_10	.11	.32	.05	.22	1.85
Region_11	.01	.07	.02	.14	.70
Region_12	.00	.06	.02	.13	.65
Region_13	.31	.46	.39	.49	.96
Student-teacher ratio	28.80	57.54	27.25	11.23	.06
Full-day	.001	.02	.01	.10	.98

Note. χ^2 statistic = 57.12; *p*-value = .000. Joint test for equality of means (rejected).

* Pairwise *t*-test for difference of means.

TABLE A3
TOURNAMENT EFFECTS, DOUBLE ROBUST METHOD

	Math		Language	
	Δ1995–96	Δ1995–97	Δ1995–96	Δ1995–97
Treat	.23	.236***	.146	.264**
	(.225)	(.084)	(.157)	(.105)
Pub	.027	−.034	−.035	−.089***
	(.043)	(.034)	(.039)	(.034)
Primary	.251	.167	.078	−.063
	(.257)	(.241)	(.186)	(.248)
Secondary	.093	.103	−.039	−.086
	(.262)	(.243)	(.191)	(.248)
College	.172	−.090	.033	−.147
	(.319)	(.277)	(.237)	(.291)
Full-day	.227	.318	.279*	.285
	(.25)	(.222)	(.161)	(.175)
Student-teacher ratio	−.002*	.001*	−.002**	−.000
	(.001)	(.001)	(.001)	(.001)
Region_2	−.103	−.050	−.126	.165*
	(.168)	(.089)	(.158)	(.098)
Region_3	−.125	−.160	−.125	−.208*
	(.149)	(.108)	(.163)	(.109)
Region_4	−.029	−.141	−.162	−.207*
	(.15)	(.106)	(.156)	(.115)
Region_5	−.159	−.092	−.231*	−.118
	(.118)	(.082)	(.134)	(.088)
Region_6	−.193	−.032	−.407***	−.098
	(.221)	(.136)	(.151)	(.121)
Region_7	.115	−.029	−.029	−.067
	(.144)	(.099)	(.159)	(.108)
Region_8	−.007	−.044	−.079	−.050
	(.154)	(.083)	(.144)	(.092)
Region_9	−.105	−.038	−.147	−.002
	(.138)	(.102)	(.148)	(.104)
Region_10	−.222	−.191*	−.248	.008
	(.171)	(.102)	(.171)	(.109)
Region_11	.314	.150	−.051	−.084
	(.265)	(.242)	(.220)	(.255)
Region_12	.061	−.065	−.309*	−.097
	(.172)	(.145)	(.175)	(.146)
Region_13	−.11	−.191**	−.248*	−.086
	(.116)	(.077)	(.131)	(.088)
Constant	−.186	−.277	.185	−.034
	(.362)	(.266)	(.279)	(.279)
Observations	1,816	1,786	1,807	1,786
R^2	.03	.03	.03	.03

Note. Robust standard errors are in parentheses. Reweighting is done using weights from Sec. III.A.
* Significant at 10%.
** Significant at 5%.
*** Significant at 1%.

TABLE A4

TOURNAMENT EFFECTS, DOUBLE ROBUST METHOD, EXCLUDING SURE LOSERS AND SURE WINNERS

	Math		Language	
	Δ1995–96	Δ1995–97	Δ1995–96	Δ1995–97
Treat	.423*	.324***	.321*	.366**
	(.218)	(.109)	(.181)	(.149)
Pub	.044	.169	.024	.085
	(.126)	(.109)	(.118)	(.113)
Primary	.478	.393	.360	.134
	(.372)	(.250)	(.278)	(.533)
Secondary	.505	.409	.255	.08
	(.406)	(.269)	(.310)	(.545)
College	1.046	.242	1.273	1.066
	(1.026)	(.374)	(.902)	(.892)
Full-day	−.272	.117	.023	.543
	(.280)	(.287)	(.239)	(.341)
Student-teacher ratio	−.005	.001	−.006**	−.003
	(.003)	(.001)	(.003)	(.002)
Region_2	−.590***	−.238	−.587***	.028
	(.172)	(.169)	(.219)	(.196)
Region_3	−.348	−.349	−.765**	−.407*
	(.389)	(.288)	(.344)	(.222)
Region_4	−.575	−.258	−.573*	−.451*
	(.413)	(.267)	(.325)	(.268)
Region_5	−.228	.007	−.330	−.128
	(.221)	(.192)	(.250)	(.201)
Region_6	−.005	.096	−.830***	−.347
	(.546)	(.351)	(.227)	(.273)
Region_7	.059	.017	−.007	.082
	(.256)	(.246)	(.276)	(.252)
Region_8	.260	−.030	−.055	−.076
	(.279)	(.186)	(.249)	(.219)
Region_9	−.761***	.060	−.732***	.006
	(.258)	(.283)	(.276)	(.213)
Region_10	−.972***	−.284	−.852***	.219
	(.257)	(.241)	(.303)	(.217)
Region_11	.228	−.078	−.179	−.006
	(.378)	(.297)	(.349)	(.562)
Region_12	−.679*	−.167	−.968***	−.062
	(.359)	(.348)	(.347)	(.269)
Region_13	−.467**	−.373**	−.538**	−.145
	(.221)	(.187)	(.241)	(.216)
Constant	−.104	−.419	.279	−.131
	(.444)	(.306)	(.379)	(.576)
Observations	637	621	628	621
R^2	.20	.17	.18	.11

Note. Robust standard errors are in parentheses. Reweighting is done using weights from Sec. III.A. Excludes schools with probability of winning less than .6 and more than .95.

* Significant at 10%.

** Significant at 5%.

*** Significant at 1%.

References

Abadie, A., and G. Imbens. 2006. "Large Sample Properties of Matching Estimators for Average Treatment Effects." *Econometrica* 74, no. 1:235–67.

Arellano, M., and S. Bond. 1991. "Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application." *Review of Economic Studies* 58:277–97.

Barlevy, G., and D. Neal. Forthcoming. "Pay for Percentile." *American Economic Review.*

Bellei, C. 2009. "Does Lengthening the School Day Increase Students' Academic Achievement? Results from a Natural Experiment in Chile." *Economics of Education Review* 28, no. 5:629–40.

Busso, M., J. DiNardo, and J. McCrary. 2009. "Finite Sample Properties of Semiparametric Estimators of Average Treatment Effects." IZA Discussion Paper no. 3998, Institute for the Study of Labor, Bonn.

Cabezas, V., J. Cuesta, and F. Gallego. 2011. "Education Outcomes in Low-Income Sectors: Evidence from Two Randomized Evaluations in Chile." Unpublished manuscript, Pontificia Universidad Católica de Chile.

Carnoy, M., I. Brodziak, A. Molina, and M. Socías. 2007. "The Limitations of Teacher Pay Incentive Programs Based on Inter-cohort Comparisons: The Case of Chile's SNED." *Education Finance and Policy* 2, no. 3:189–227.

Contreras, D., O. Larrañaga, L. Flores, F. Lobato, and V. Macias. 2005. "Políticas educacionales en Chile: Vouchers, concentración, incentivos y rendimiento." In *Uso e impacto de la información educativa en América Latina*, ed. Santiago Cueto, 61–110. Santiago: PREAL.

Contreras, D., P. Sepulveda, and S. Bustos. 2010. "When Schools Are the Ones That Choose: The Effects of Screening in Chile." *Social Science Quarterly* 91, no. 5: 1349–68.

Figlio, D., and L. Kenny. 2007. "Individual Teacher Incentives and Student Performance." *Journal of Public Economics* 91:901–14.

Fryer, R. 2011. "Teacher Incentives and Student Achievement: Evidence from New York City Public Schools." Working Paper no. 16850, National Bureau of Economic Research, Cambridge, MA.

Gallego, F. 2008. "Efectos del SNED en resultados del proceso educativo." Unpublished manuscript, Catholic University of Chile.

Glewwe, P., N. Ilias, and M. Kremer. 2010. "Teacher Incentives." *American Economic Journal: Applied Economics* 2, no. 3:205–27.

Goodman, S., and L. Turner. 2010. "Teacher Incentive Pay and Educational Outcomes: Evidence from the New York City Bonus Program." Unpublished manuscript, Columbia University.

Hanushek, E., J. Kain, and S. Rivkin. 1999. "Do Higher Salaries Buy Better Teachers?" Working Paper no. 7082, National Bureau of Economic Research, Cambridge, MA.

Harbinson, R., and E. Hanushek. 1992. *Educational Performance of the Poor: Lessons from Rural Northeast Brazil.* New York: Oxford University Press.

Koretz, D. M. 2002. "Limitations in the Use of Achievement Tests as Measures of Educators' Productivity." *Journal of Human Resources* 37, no. 4:752–77.

Lavy, V. 2002. "Evaluating the Effect of Teachers' Group Performance Incentives on Pupil Achievement." *Journal of Political Economy* 110, no. 6:1286–1317.

Mizala, A., and P. Romaguera. 2005. "Teachers' Salary Structure and Incentives in Chile." In *Incentives to Improve Teaching: Lessons from Latin America*, ed. E. Vegas. Washington, DC: World Bank Press.

Mizala, A., and M. Urquiola. 2007. "School Markets: The Impact of Information Approximating Schools' Effectiveness." Working Paper no. 13676, National Bureau of Economic Research, Cambridge, MA.

Muralidharan, K., and V. Sundararaman. 2010. "The Impact of Diagnostic Feedback to Teachers on Student Learning: Experimental Evidence from India." *Economic Journal* 120, no. 546:F187–F203.

———. 2011. "Teacher Performance Pay: Experimental Evidence from India." *Journal of Political Economy* 119, no. 1:39–77.

Neal, D., and D. W. Schanzenbach. 2010. "Left Behind by Design: Proficiency Counts and Test-Based Accountability." *Review of Economics and Statistics* 92, no. 2:263–83.

Robins, J. M., and A. Rotnitzky. 1995. "Semiparametric Efficiency in Multivariate Regression Models with Missing Data." *Journal of the American Statistical Association* 90, no. 429:122–29.

SIMCE. 1989–99. "Base de datos SIMCE (1989–1999)." Santiago, Chile: SIMCE, Ministerio de Educación.

Springer, M. G., D. Ballou, L. Hamilton, V. Le, J. R. Lockwood, and D. F. McCaffrey. 2010. "Optimal Incentives for Public Sector Workers: The Case of Teacher-Designed Incentive Pay in Texas." National Center on Performance Incentives, Nashville, http://www.performanceincentive.org.

Urquiola, M., and E. Verhoogen. 2009. "Class-Size Caps, Sorting, and the Regression-Discontinuity Design." *American Economic Review* 99, no. 1:179–215.