ELSEVIER

Journal of Development Economics 84 (2007) 61 – 75

JOURNAL OF Development ECONOMICS

www.elsevier.com/locate/econbase

Socioeconomic status or noise? Tradeoffs in the generation of school quality information ☆

Alejandra Mizala [a], Pilar Romaguera [a], Miguel Urquiola [b,*]

[a] *Universidad de Chile, Chile*
[b] *Columbia University, USA*

Received 14 March 2006; received in revised form 7 September 2006; accepted 11 September 2006

Abstract

This paper calculates a time series of simple, standard measures of schools' relative performance. These are drawn from a 1997–2004 panel of Chilean schools, using individual-level information on test scores and student characteristics for each year. The results suggest there is a stark tradeoff in the extent to which rankings generated using these measures: i) can be shown to be very similar to rankings based purely on students' socioeconomic status, and ii) are very volatile from year to year. At least in Chile, therefore, producing a meaningful ranking of schools that may inform parents and policymakers may be harder than is commonly assumed.

JEL classification: I2; O1
Keywords: Tradeoff; Socioeconomic status; School performance

1. Introduction

Improving public service delivery, particularly as it affects the poor, has come to be seen as one of the central challenges in development policy. In the case of education, this reflects that service quality is extremely low, in some cases due to basic problems like high absenteeism (Chaudhury et al., 2006) or capture (Reinikka and Svensson, 2005a). Even in settings in which such issues

☆ We thank Samuel Berlinski, Jishnu Das, Esther Duflo, Marcelo Henriquez, Leigh Linden, Patrick McEwan, Jonah Rockoff and Timothy Vogelsang for useful conversations and comments. We are also very grateful to the SIMCE office at Chile's Ministry of Education for providing data. All remaining errors are ours only.
* Corresponding author.
E-mail address: msu2101@columbia.edu (M. Urquiola).

0304-3878/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.jdeveco.2006.09.003

have been addressed, service quality and outcomes need improvement. Pritchett (2004), for instance, points out that outside of East Asia, developing country performance on standardized tests is frequently dismal.

At the same time, there is no consensus on how best to go about improving service quality. One approach emphasizes external control and incentives. Banerjee and Duflo (2006) review how these have been used to address teacher absence,[1] and as we discuss below, initiatives in several countries have sought to tie rewards to student performance.

Another approach emphasizes providing users with greater information, such that they can better choose and monitor providers. In the words of the 2004 World Development Report (World Bank, 2004), "increasing poor clients' choice and participation in service delivery will help them monitor and discipline providers. Raising poor citizens' voice, through the ballot box and widely available information, can increase their influence with policy makers…" There is not much reliable evidence on the effects of information provision per se. While Banerjee and Duflo (2006) point out that local monitoring by itself may not be that effective at reducing absenteeism,[2] Reinikka and Svensson (2005b) find that a newspaper campaign which helped parents and head teachers monitor fiscal transfers reduced the diversion of funds. Das et al. (2006), describe a project that by randomizing the provision of school report cards across school markets in Pakistan, should eventually provide valuable information.

In this paper, we analyze these issues in the context of Chile, a country which has taken the provision of educational choice, incentives, and information very seriously. As Chile has found, whichever of these approaches one wants to emphasize, a crucial input is an assessment of schools' performance, including in many cases, a ranking of institutions that can be used to inform parents or allocate rewards or penalties in accountability-type schemes.[3] Additionally, in view of the sparse knowledge on the causal impact of school inputs, economists generally consider test-based rankings, because they judge schools on a key output, as preferable to input-based measures.

In this paper we argue that producing such information might be harder than it seems, even in a country like Chile, which has more and better data than the modal developing country. To make this case, we first note that any attempt to construct a test-based school ranking faces two challenges. The first reflects that students are not randomly assigned to schools, and some institutions may therefore perform better because they enroll "better" children, rather than because they are inherently more productive. This issue, while difficult to address, is well understood.

The second challenge arises because schools' mean test scores can provide a "noisy" measure of performance-transitory factors might determine that schools that do relatively well one year have a systematic tendency to do relatively poor the next, even if their underlying productivity

[1] Specifically, Duflo and Hanna (2005) present a randomized experiment in which teacher presence was monitored using cameras, and in which bonuses were paid for good attendance. This approach seems to have been effective, in contrast to a situation where authority and payments were controlled by headmasters, as discussed in Kremer and Chen (2001).

[2] See Banerjee et al. (2004) and Kremer and Vermeesh (2005).

[3] While Chile has emphasized school choice more than most countries, it is far from alone in implementing accountability efforts. As of 2004, 16 states in the U.S. used rankings to allocate rewards to high-performing schools; 36 provided assistance to low-performing ones; and 27 administered sanctions to low-performing schools (Skinner and Staresina, 2004). Other countries have experimented with rewarding schools or teachers based on the test performance of their students, such as Israel (Lavy, 2002), Kenya (Glewwe et al., 2003), Mexico (McEwan and Santibañez, 2004), and Chile, the subject of this paper.

remains stable. In such a case, rankings will display substantial volatility and could easily mislead parents and policy makers.[4]

This paper's contribution is to suggest that these two problems may be linked such that attempts to alleviate one substantially aggravate the other, leaving open the possibility that producing a meaningful ranking of schools is harder than may appear. Making this case is difficult because it is ultimately impossible to credibly identify what component of each school's performance is due to its own value added, and what components might be due to its students' background or to transitory factors unrelated to its real productivity.

Instead, the approach we take here is to make two assumptions. First, we assume that policy makers would worry about measures which produce rankings that can be shown to be very similar to those that would result from simply ordering schools based on their students' socioeconomic characteristics. While such orderings might indeed reflect schools' true productivity, one would worry, for instance, about accountability schemes that penalized schools simply for taking poor children. Second, we assume that policy makers would also worry if the rankings produced by any given measure displayed high year to year volatility, in the extreme, producing accountability-based rewards that mimicked a lottery. Again, while volatility might reflect true changes in schools' underlying productivity, one would be concerned about policies that led households to switch schools too often,[5] or about measures that struck teachers as unfair or blunted the incentives they faced.

With this background, we calculate several simple performance measures using Chile's SIMCE national standardized testing system, which since the mid-1990s has collected information on students' characteristics and their performance. Specifically, we obtain a time series of observations on schools' average scores, schools' average scores adjusted to remove individual background differences, residuals from such regressions at the individual and school level, and year to year changes in scores. Thus, the set of measures one can use to rank schools in Chile is more complete than that which would be feasible in the average developing country.

Using these, a first finding is that rankings based on average school test score levels are essentially equivalent to rankings based on schools' average socioeconomic status (henceforth SES). For example, an ordering of schools based on their average test score is very close to one based on their average mothers' schooling. A second result is that a ranking based on regression-adjusted levels hardly differs from one generated using simple levels. This might reflect, among other factors, the substantial stratification observed in the Chilean school system. It also implies that in order to generate findings that do not by and large reflect SES, one is forced to use residuals and changes in school average scores from year to year. Our third finding is that these measures, particularly the latter, are very volatile, confirming results in Kane and Staiger (2002).

Taken together, these findings suggest that at least in Chile, producing useful rankings of schools is difficult. The simplest ones essentially reflect SES, and moving towards measures that are less explained by student background results in rapid increases in volatility—an undesirable tradeoff under our assumptions.

Although we do not know to what extent these findings generalize to other settings, one implication is that while using information to improve service quality along dimensions like

[4] This issue is highlighted by Kane and Staiger (2001) and Chay et al. (2005).

[5] Hanushek et al. (2004) suggest that turnover can entail significant negative externalities.

absenteeism might be relatively simple, doing so to improve educational quality more broadly may be more challenging.

Our findings also imply Chile might do well to invest more in data collection and analysis in this area. First, following some states in the U.S., it could gather data that would allow calculating individual-level gains as students progress through the school system. Second, it could explore the use of the time series dimension in performance measures to produce filtered estimates in the manner discussed by Kane and Staiger (2002). Unfortunately, the literature on how to best implement these techniques is limited, and they are not in widespread use even in the U.S. Neither of these initiatives would provide complete solutions to the problems raised here, but both might help improve the quality and usefulness of school comparisons.

The rest of the paper is organized as follows. Section 2 presents some background and describes the data. Section 3 presents a framework and Section 4 describes the results. Section 5 concludes.

2. Background and data

Chile has been at the forefront of efforts to use choice, incentives, and information to improve service delivery in the educational sector. First, in 1981 it introduced an unrestricted voucher system, which allows any child who wants a voucher to use it at almost any public or private school.[6] At this time, mechanisms to provide information that might aid parental choice entered policy discussions. This contributed to the creation of the PER[7] testing program, which first collected data in 1982. In the event, this program did not prosper and was stopped by 1984; to our knowledge its results were not used or publicized.

In 1988 a new testing system, SIMCE,[8] came into existence, and as of the mid-1990s, its results had been widely disseminated, partially by way of listings of individual schools' performance in major newspapers. The government also began using SIMCE scores to allocate resources. For instance, in 1990 the *P-900* program began using the mean of 4th grade test scores to allocate aid to about 900 under-performing schools.[9]

Additionally the government began to consider using test scores to promote accountability and transmit incentives. In part, this was meant to address the fact that in smaller markets (e.g. rural areas with few schools) competition might be limited. In 1996 the SNED[10] system began financially rewarding the teachers in schools that perform well, choosing these from within predetermined "homogenous groups" of schools in each of 13 administrative regions.[11]

[6] About 90% of Chilean children use this subsidy to attend public or private schools, where the latter include religious and for-profit institutions. For further description of the voucher system and the private school sector, see Mizala and Romaguera (2000) and Urquiola and Verhoogen (2006).

[7] *Programa de Evaluación del Rendimiento Escolar.*

[8] SIMCE stands for *Sistema Nacional de Medición de la Calidad de la Educación*, and employs an Item Response Theory methodology.

[9] Chay et al. (2005) describe this program and explore how mean-reverting noise complicates its allocation and evaluation. P-900 ended about 10 years after its inception, but has been replaced by other targeted programs allocated at least partially as a function of schools' performance.

[10] *Sistema Nacional de Evaluación del Desempeño de los Establecimientos Educativos Subvencionados*.

[11] These are constructed stratifying institutions according to their location (urban and rural), the educational level at which they operate (primary or secondary), and their students' SES. The SNED index relies mainly on test score levels and changes in levels. Nonetheless, 35% of its value is based on non-test related variables, like drop out rates and parental participation/perceptions. For a more thorough description of the SNED and a discussion of its impact, see Mizala and Romaguera (2002, 2004).

In this paper we use SIMCE information, which has among its strengths:

1) The time series it yields is fairly long for this type of information, since there is a total of eleven years of individual-level data (1993–2003) for each school.[12]
2) It provides essentially a census of schools, covering both public and private institutions.
3) For eight of the eleven years (1997–2004) with individual level data, there are also detailed individual level controls collected via a parental questionnaire. These yield information on students' gender, their mothers' and fathers' schooling, and their household income.[13]

Of course, these data are not without disadvantages. One of the key ones is that while the SIMCE provides a panel of schools, it does not track students over time, so that it does not permit the calculation of individual-level changes in performance (*gain scores*, in the terminology of Kane and Staiger, 2002). Further, only one grade is visited each year, alternating between the 4th, 8th, and 10th. Thus, calculating year to year *changes* in test scores requires comparing different grade levels, which requires us to focus on relative comparisons in most analyses.

Out of roughly 5000 schools providing primary and secondary education, we restrict attention to four samples for which we have a panel:

1) A group of 701 schools with a score each year between 1997 and 2004—those in which individual level controls are available. A drawback with this sample is that it requires comparing schools' relative performance in the 4th, 8th, and 10th grades. This implies that it is composed of relatively large schools, since these are more likely to offer all three grades—a relevant fact because school size is an important determinant of ranking volatility.
2) A sample of 3331 schools that have a score in the five years between 1997 and 2004 in which either a 4th or an 8th grade score was collected (1997, 1999, 2000, 2002, and 2004). While it still has individual level controls, this sample is larger because it incorporates schools that offer only primary instruction (comprised of grades 1–8 in Chile), which as elsewhere tend to be significantly smaller and more numerous.[14]
3) A sample of 3840 schools that have 8th grade scores in 1993, 1995, 1997, 2000, and 2004. This sample no longer allows us to control for individual characteristics (which are not available for the two first years), but it does permit comparisons to concern a single grade.
4) A sample of 1414 schools with a 10th grade score for 1998, 2001, and 2003. These are the years in which the Ministry of Education indicates test scores are comparable over time. Thus in, this sample we can compare "raw" rather than relative performance.

For reasons of space, we present results on only the first of these samples. We discuss those from the others (which generate similar conclusions), and they are available upon request.

3. Framework

In order to understand what different test-based performance measures capture, consider a framework explaining test score determination. Let y_{ijt} denote the score for individual i in school

[12] The series actually starts in 1988 (with 1989 excluded) if one uses aggregated school level data.
[13] Some years' databases include more information on parental job characteristics and household assets.
[14] Additionally, as in most developing countries, the enrollment rate for primary education in Chile is higher, and had already approached 100% well before 1997.

j and year t, and suppose the vector X_{ijt} contains socioeconomic status (SES) information. Assume that students' scores are given by:

$$y_{ijt} = \delta_{jt} + X'_{ijt}\beta + u_{ijt}$$

where u_{ijt} is a mean zero iid error, and δ_{jt}, a school effect, can be decomposed as

$$\delta_{jt} = \tilde{\delta}_j + z_{jt} + \varepsilon_{jt}$$

where $\tilde{\delta}_j$ does not vary over time, and ε_{jt} is a mean zero iid time series. Suppose also that z_{jt} is an autoregressive process with one lag:

$$z_{jt} = \rho z_{j,t-1} + v_{jt}$$

and assume that v_{jt} is a mean zero iid time series, and that ε_{jt} and v_{jt} are uncorrelated with each other, so that z_{jt} and ε_{jt} are orthogonal—z_{jt} represents the persistent component and ε_{jt} represents a transitory component.

In sum, students' test scores at time t are given by:

$$y_{ijt} = \tilde{\delta}_j + \rho z_{j,t-1} + v_{jt} + \varepsilon_{jt} + X'_{ijt}\beta + u_{ijt} \quad (1)$$

Ideally, at any point in time t we would like to rank schools according to $\tilde{\delta}_j + \rho z_{j,t-1}$, the portion of students' achievement that is due to the school and is not transitory.

4. Results

Using this framework, consider five simple test-based performance measures, all but one of which rely only on cross-sections of data. Simplicity is often considered desirable for accountability measures, since it is useful if parents and teachers are able to understand why their school did or did not get selected for an award or penalty. Perhaps because of this, the measures we consider (or variants thereof) are already explicitly or implicitly in use in Chilean accountability-related programs, and they are also common in the U.S.

4.1. Levels (mean scores)

To begin, consider the simplest situation, in which only one cross section of student-level data is used, and schools are ranked based on their mean scores. As Hanushek (2004) states, this is the most basic measure, and makes an appearance in essentially all accountability schemes. In a regression setting, one can drop the t subscript from Eq. (1) and implement:

$$y_{ij} = \mathbf{a} + u_{ij} \quad (2)$$

where $\mathbf{a}$ is vector of school-specific dummies used to rank schools. In each year we normalize test scores to have a mean of zero and a standard deviation of one. Each school's intercept therefore indicates its students' average relative position in the distribution of scores.

Panel A in Table 1 summarizes these regressions for each cross-section of data considered. For instance, for 1997, column 1 presents a regression of 47,350 student level language test scores on 701 school dummies.[15] A relevant result is that these regressions have R^2's generally between 0.3

[15] We omit analogous results for math, which produce very similar conclusions, for the sake of space. Math results for Tables 1–4 are available from the authors upon request. We also note that we have replicated these tables in urban and rural samples, and within the largest municipalities, with similar conclusions.

Table 1
Regressions using individual level data for the 701 school sample

Dependent variable: language score	1997 (8th grade)	1998 (10th grade)	1999 (4th grade)	2000 (8th grade)	2001 (10th grade)	2002 (4th grade)	2003 (10th grade)	2004 (8th grade)
Panel A: School dummies only								
701 School dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	47,350	25,356	51,865	50,288	50,127	44,846	57,420	50,830
R^2	0.353	0.473	0.337	0.309	0.376	0.320	0.389	0.301
Panel B: Individual controls only								
Dummy for sex	Yes	Yes	–	Yes	Yes	Yes	Yes	Yes
No. of mothers' schooling dummies	24	23	23	49	63	99	41	47
No. of fathers' schooling dummies[a]	24	23	23	49	63	99	41	47
No. of household income dummies	24	24	24	14	24	14	17	16
N	47,350	25,356	51,865	50,288	50,127	44,846	57,420	50,830
R^2	0.196	0.246	0.236	0.191	0.243	0.214	0.259	0.208
Panel C: School dummies and individual controls								
701 School dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Dummy for sex	Yes	Yes	–	Yes	Yes	Yes	Yes	Yes
No. of mothers' schooling dummies	24	23	23	49	63	99	41	47
No. of fathers' schooling dummies[a]	24	23	23	49	63	99	41	47
No. of household income dummies	24	24	24	14	24	14	17	16
N	47,350	25,356	51,865	50,288	50,127	44,846	57,420	50,830
R^2	0.373	0.486	0.358	0.332	0.393	0.344	0.406	0.328

Note: Authors' calculations using SIMCE data. The 701 schools covered are those that had valid test scores in every single cross-section, 1997–2004. Panel A contains regressions where the only independent variables are a full set of school dummies. Panel B contains regressions on the control variables only, and Panel C combines both school dummies and controls. For the most flexible specification, the controls are dummies indicating all the possible responses parents could have given to questions regarding their schooling and household income.

[a] In 1997, the question asked about the household head's (rather than the father's) schooling.

Table 2
School level regressions for the 701 school sample

Dep. variable: Language score	1997 (8th grade)	1998 (10th grade)	1999 (4th grade)	2000 (8th grade)	2001 (10th grade)	2002 (4th grade)	2003 (10th grade)	2004 (8th grade)
Dummy for sex	Yes	Yes	–	Yes	Yes	Yes	Yes	Yes
No. of mother's ed. dummies	24	23	23	49	63	99	41	47
No. of father's ed. dummies [a]	24	23	23	49	63	99	41	47
No. of household income dummies	24	24	24	14	24	14	17	16
N	701	701	701	701	701	701	701	701
R^2	0.754	0.797	0.826	0.776	0.857	0.831	0.853	0.831

Note: Authors' calculations using SIMCE individual-level data aggregated to the school level. The 701 schools covered are those that had valid test scores in every single cross-section, 1997–2004.
[a] In 1997, the question asked about the household head's (rather than the father's) schooling.

and 0.4. As Kane and Staiger emphasize, this suggests that only about one third of the variation in test scores is between schools. Put otherwise, in many markets the worst student at a "good" school will score worse than the best students at a "bad" one.

In terms of the two challenges described above, the ranking implicit in Eq. (2) is of course extreme in making no attempt to control for the fact that some schools might rank higher simply because they enroll higher SES children; in terms of Eq. (1) the ordering generated by **a** explicitly includes the contribution of X_{ijt}.

Table 2 shows that this is indeed a concern by regressing the components of vector **a** on school-level average observable SES—this amounts to a regression of average test scores on average SES at the school level (the sample size is thus 701 schools, as opposed to about 50 thousand students). As the table shows, depending on the year, between 75 and 86% of the variation in test scores can be explained by parental schooling and household income. This implies that a ranking based on school-level average test scores levels is not far from one based on which schools enroll "better" children.

For example, consider a program that selects the top fifth of schools, and consider the selection to such program that would be produced by: i) ranking schools according to their mean language score, and ii) ranking schools according to average mothers' schooling. Nationwide, these two measures would agree on the selection or non-selection of about 85% of all schools.[16] In short, the possibility that rankings might simply reflect SES is a distinct one regarding test score levels.

For an initial glimpse at how important the second concern, the volatility in rankings, might be with this measure, consider to what extent the rankings it produced might resemble a lottery over time, as one would expect if average test scores are heavily influenced by onetime, mean-reverting shocks. Table 3 follows Kane and Staiger (2001) and presents the results of this exercise for language scores. Panel A refers to a hypothetical program that selects the top 20% of schools, and column 1 contains the outcome that one would observe if schools' performance were totally stable: 80% of schools would never appear in the top quintile, and 20% would be in this category all eight years. At the opposite extreme, column 2 supplies the frequencies that would be observed if the program were essentially a lottery—if each school had an independent 0.2 probability of being selected each year. In this case one would expect that over the eight years considered, only 17% of schools never get selected. The modal school would end up in the top quintile once, and

[16] A greater correspondence is possible if one also uses information on the fathers' schooling and household income. A similar correspondence would result for a program that selected the bottom fifth of performers.

Table 3
Comparison of the frequency in top or bottom 20% produced by different rankings—701 school sample, Language

	Levels, adjusted levels, and residuals						Changes		
	Certainty	Lottery	Levels	Adjusted levels	Residuals, individual data	Residuals, school level data	Certainty	Lottery	Gains
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A: Number of times schools appears in the top 20%									
Never	80.0	16.7	59.3	58.4	45.7	32.5	80.0	21.0	16.6
1 year	0.0	33.6	9.1	8.1	18.0	24.7	0.0	36.7	40.0
2 years	0.0	29.4	5.0	6.4	11.3	17.7	0.0	27.5	31.7
3 years	0.0	14.7	5.7	6.4	7.1	11.3	0.0	11.5	10.7
4 years	0.0	4.6	5.3	4.9	4.1	7.7	0.0	2.9	1.4
5 years	0.0	0.9	3.0	4.6	6.4	3.3	0.0	0.4	0.0
6 years	0.0	0.0	3.9	4.1	3.4	1.0	0.0	0.0	0.0
7 years	0.0	0.0	4.1	3.4	2.1	1.0	20.0	0.0	0.0
All 8 years	20.0	0.0	4.6	3.7	1.9	0.9	–	–	–
Panel B: Number of times schools appears in the bottom 20%									
Never	80.0	16.7	68.1	65.5	53.8	34.5	80.0	21.0	19.1
1 year	0.0	33.6	5.1	6.9	12.4	22.5	0.0	36.7	34.7
2 years	0.0	29.4	2.3	2.3	8.4	17.3	0.0	27.5	33.5
3 years	0.0	14.7	1.4	2.1	6.1	10.0	0.0	11.5	11.7
4 years	0.0	4.6	3.9	3.9	4.4	8.3	0.0	2.9	1.0
5 years	0.0	0.9	3.9	5.6	3.4	4.6	0.0	0.4	0.0
6 years	0.0	0.0	2.6	2.9	4.1	2.0	0.0	0.0	0.0
7 years	0.0	0.0	5.9	5.1	4.6	0.9	20.0	0.0	0.0
All 8 years	20.0	0.0	7.0	5.9	2.7	0.0	–	–	–
Panel C: Percentage ever in top and bottom 20%	–	–	0.3	0.9	10.6	35.8	–	–	72.5

Note: Panel A refers to a hypothetical program that selects the top 20% of performers under each measure and scenario. Panel B presents an analogous exercise when selection is for the bottom 20%. Panel C indicates the percentage of schools that over the whole period (1997–2004), would have ever appeared in both the top *and* the bottom quintile under each measure. Columns 1 and 7 (for levels and gains, respectively) are benchmarks describing the distribution of selections if performance were completely stable. Columns 2 and 8 are benchmarks describing the distribution if selection was essentially based on a lottery.

one would expect no school would be selected every year. Of course, while no useful measure would produce total stability, we assume one would worry about measures that resembled a lottery. Panel B presents a similar exercise selecting the bottom 20% of schools.

Column 3 summarizes what happens with the rankings that emerge using levels (**a** in regression 2). The observed distribution is closer to that one would expect if schools' performance were stable. This is not surprising to the extent that this measure explicitly incorporates SES, and to the extent that one expects schools' position in the relative SES distribution to be stable over time. Additionally, in the framework of Eq. (1), this measure also incorporates δ, which by definition does not vary across periods.

In short, under the assumptions made above, if one thinks of a tradeoff in terms of the extent to which measures reflect (observable) SES, and the extent to which they display "excess" volatility, school-level mean test scores clearly suffer much more from the former.

4.2. Adjusted levels

An alternative measure is obtained by regression-adjusting scores to remove the influence of individual background characteristics. One would run

$$y_{ij} = \mathbf{b} + X'_{ij}\beta + u_{ij} \tag{3}$$

and use **b**, the vector of school-specific intercepts, to rank schools. This is the approach that Kane and Staiger (2001) implement to generate annual school performance measures, which they label *adjusted levels*. It can be calculated in Chile, where in many years we know students' gender, their mothers' and fathers' years of schooling, and their household income.

For background on the impact of including controls, Panel B in Table 1 presents regressions of students' scores on only these variables. That is, this panel summarizes a regression:

$$y_{ij} = \alpha + X'_{ij}\beta + u_{ij},$$

where α is a constant. For instance, column 1 regresses 47,350 students' language scores on 72 dummies that characterize SES.[17] The R^2 in these regressions is generally about 0.2. In short, students' observable characteristics explain a relatively small but certainly non-trivial proportion of the variation in their test scores.

Panel C adds the school effect, summarizing the results of regressions like Eq. (3) for every available cross-section. Perhaps the most interesting result here is that within each cross-section, the fit in panels A and C is very similar. Put otherwise, adding SES controls to a regression that already includes school dummies changes the specifications' ability to explain test scores very little, despite the fact that on their own these controls explain a nontrivial part of the variation in scores (Panel B). For instance, in column 1 going from specification (2) to specification (3) increases the R^2 by only about two percentage points, from 0.35 to 0.37 (despite the fact that by themselves the controls produce an R^2 of 0.2).

Perhaps surprisingly, therefore, these results suggest that in Chile, school rankings calculated from Eqs. (2) and (3) are very similar—in the terminology of Kane and Staiger (2001), *adjusted levels* might as well be *levels*. Indeed schools' coefficients from **a** and **b** are highly correlated: the lowest within year correlation coefficient for these is 0.988. If they were used in a hypothetical program that selected the top 20% of schools, the allocations based on them would agree on about 95% of all schools.

The strength of this surprising result might be specific to Chile, and may be due to a variety of factors. Among them might be the extensive sorting observed in Chile's school system, in part facilitated by substantial school choice.[18] In practical terms, the close correspondence between **a** and **b** suggests that the system displays enough stratification such that the fact that a given student is enrolled in one or another school transmits a lot of information about her observable SES, and importantly, perhaps about other unobserved characteristics that are important determinants of achievement.

One way of checking this for observables, is by confirming that schools are indeed more homogeneous in SES than in test scores. Table 4 does so by presenting regressions of language

[17] The number of dummies varies from year to year because of changes in the precise nature of the questions asked, and in the number of options or ranges respondents could choose from. Instead of the 72 dummies we could just include three variables: mother's schooling, father's schooling, and household income. We opt for the dummy specification to allow for the greatest flexibility. In the event, the R^2s produced by the two approaches are quite similar.

[18] See for instance Hsieh and Urquiola (2006).

Table 4
Individual level regressions for the 701 school sample—Language

	1997 (8th grade)	1998 (10th grade)	1999 (4th grade)	2000 (8th grade)	2001 (10th grade)	2002 (4th grade)	2003 (10th grade)	2004 (8th grade)
Panel A — Language score								
701 School dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	38,656	18,559	44,205	36,410	31,532	35,193	42,701	28,841
R^2	0.346	0.460	0.328	0.299	0.374	0.317	0.390	0.318
Panel B — Mothers' schooling								
701 School dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	38,656	18,559	44,205	36,410	31,532	35,193	42,701	28,841
R^2	0.453	0.465	0.459	0.439	0.397	0.445	0.471	0.480
Panel C — Fathers' schooling								
701 School dummies		Yes	Yes	Yes	Yes	Yes	Yes	Yes
N		18,559	44,205	36,410	31,532	35,193	42,701	28,841
R^2		0.483	0.494	0.469	0.429	0.460	0.500	0.500
Panel D —Household income								
701 School dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	38,656	18,559	44,205	36,410	31,532	35,193	42,701	28,841
R^2	0.609	0.693	0.725	0.683	0.707	0.720	0.701	0.706
Panel E — ln(household income)								
701 School dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	38,656	18,559	44,205	36,410	31,532	35,193	42,701	28,841
R^2	0.504	0.647	0.659	0.620	0.637	0.660	0.654	0.640

Note: For each cross-section, the columns present regressions of the dependent variable on a full set of school dummies.

scores, mothers' schooling, fathers' schooling, and household income on a full set of school dummies. On average, the R^2 on test scores is about 0.3. The R^2 on parental schooling is generally about ten percentage points higher each year, and that on income is about 35 percentage points higher—in other words, a full set of school dummies explains almost 70% of the variation in household income (and about two thirds if the dependent variable is the log of income, as in Panel E).

In terms of the stability of rankings, columns 3 and 4 in Table 3 suggest that this measure not surprisingly produces a distribution very similar to a simple ranking based on schools' mean scores, one that is (loosely) not far from that one would expect if schools' relative performance were close to stable. In short, by the criteria we consider here, measures based on mean test scores and mean adjusted scores share key advantages and disadvantages.

4.3. Individual level residuals

A third measure, one frequently mentioned in educational research, is obtained by focusing on residuals, selecting schools whose students perform better than one would predict given their SES. Specifically, we run a regression with controls and a single constant

$$y_{ij} = \alpha + X'_{ij}\beta + u_{ij}, \tag{4}$$

and rank schools according to their average residual, $\overline{u}_{ij}$.

This measure is, partially by design, less influenced by observable SES. On the other hand, one might expect it to display more volatility because it no longer contains a school-specific effect, which as the comparison of regressions (2) and (3) suggests, is very good at absorbing determinants of test scores. Table 3, column 5, shows that the ranking obtained from Eq. (4) is indeed more volatile—the distribution of selections over time moves significantly closer to that of a hypothetical lottery. The last row shows that under this measure, over the eight years we observe, roughly 1 in 10 schools would have been singled out *both* as exceptionally good (top 20%) and exceptionally bad (bottom 20%).

4.4. Residuals at the school level

Another possibility is to take residuals from a school-level regression. One would run

$$\bar{y}_j = \alpha + \bar{X}_j\beta + e_j, \tag{5}$$

and rank institutions according to the residuals e. Table 2 first summarized the results for such specifications in each year. One salient finding was that the R^2s from such school level regressions are substantially higher than those from individual level specifications. In the individual level regressions, this statistic hovered around 0.2 depending on the year, while in the school-level regressions it is closer to 0.8.

The improved fit as one moves from individual to school level observations might be due to a series of factors. First, it might simply be that measurement error in SES is mitigated by aggregation. Second, perhaps the process by which students and schools are matched results in substantial sorting by household SES, but that households have children of different ability. As stated above, there could also be significant sorting by schools on unobserved determinants of achievement, such as motivation. Any or a combination of these mechanisms would contribute to explaining the somewhat paradoxical result that controlling for individual level SES has almost no effect on school rankings, yet once aggregated to the school level, SES variables explain up to four fifths of the variation in test scores.

In any case, the high R^2 at this level might mean that using the ranking implicit in Eq. (5) is a promising way to control for SES, since it nets out many observable components. As Table 3 (column 6) illustrates, however, this significantly increases concerns about the volatility of rankings. The distribution produced moves even closer to a lottery. Over the eight years of data we have, using this measure would have resulted in more than 65% of schools making the top quintile at some point. Further, 36% would have made an appearance in *both* the best and the worst performing groups—by the time we reach this measure, one would worry year to year rankings might generate some confusion.

4.5. Changes

A final possibility we consider, one often cited, is to rank schools according to the difference in their average scores from year to year, which in some sense may capture value added. Perhaps the best measure in this regard is what Kane and Staiger (2002) term *gain scores*, which requires tracking individual students' progress. While this is not feasible in Chile, we can compare how schools' relative positions change between years, with the substantial caveat that in the present (701 school) sample, this requires comparing schools' scores at different grade levels, since a given grade is not visited in consecutive years.

In any case, in terms of the two criteria described above, gain scores would seem particularly attractive as a way to control for SES, since if schools are stable in the type of children they serve, their average SES characteristics, including those unobserved, will be differenced out from year to year. Consider periods $t=0$ and $t=1$, with average test scores:

$$\bar{y}_{j0} = \bar{\delta}_j + z_{j0} + v_{j0} + \varepsilon_{j0} + \bar{X}'_{j0}\beta + \bar{\mu}_{j0}$$

$$\bar{y}_{j1} = \bar{\delta}_j + \rho z_{j0} + v_{j1} + \varepsilon_{j1} + \bar{X}'_{j1}\beta + \bar{\mu}_{j1}$$

Assuming that $\bar{X}_{j0}=\bar{X}_{j1}$ (i.e., assuming schools' socioeconomic composition is stable), the change eliminates the influence of SES, and is given by

$$\Delta = y_{j1}-y_{j0} = (\rho-1)z_0 + (v_{j1}-v_{j0}) + (\varepsilon_{j1}-\varepsilon_{j0})$$

This may come at the cost of higher volatility, however, if $(v_{j1}-v_{j0})+(\varepsilon_{j1}-\varepsilon_{j0})$ is large relative to $(\rho-1)z_{j0}$ (particularly since taking changes also eliminates $\bar{\delta}$). As column 9 in Table 3 shows, the distribution in this case indeed is very close to that which a lottery would generate. Over eight years (and seven calculated changes), more than 80% of schools would have been in the top group at some point, and more than 70% would have been in *both* the best and worst group.

4.6. Results with alternate samples

In results available upon request, we first replicated the results in Tables 1–4 for a sample of 3331 schools that form a panel with a valid 4th or 8th grade score in every year in which these were collected. In this case we have five rather than eight years of data, and only four observed gain scores. The results and conclusions that emerge from this group, are in line with those for the smaller sample.

We also addressed the possibility that the volatility in changes (column 9, Table 3) is driven by the fact that we consider different grades across years, which could happen, for instance, if schools' performance is very heterogeneous at different grade levels. We therefore consider a sample of 3840 schools that had an 8th grade score in 1993, 1995, 1997, 2000, and 2004. While this allows us to focus on only one grade, the years considered are obviously no longer consecutive, and the inclusion of 1993 and 1995 means that we cannot use individual level controls. The key results (also available from the authors), however, suggest that volatility of the change measures is just as marked in this case.

Finally, for 1998, 2001, and 2003 we use 10th grade scores which the Ministry of Education indicates are comparable over time. Thus in this sample we can compare the schools' "raw" rather than relative performance, and the conclusions are similar again.

5. Discussion and conclusions

A meaningful ranking of schools would be useful in efforts to improve educational service delivery—from transmitting incentives to teachers and principals, to enhancing parental school choice. This paper suggests, however, that at least in Chile this might be harder to produce than is commonly thought. Specifically, using several cross sections to calculate commonly-used school performance measures, we find there is a clear tradeoff in the extent to which these generate rankings that: i) essentially reflect students' SES, and ii) display large year-to-year volatility. This is an undesirable tradeoff under the two assumptions we made: that policymakers would worry

about measures that largely reflect SES, and that they would worry about measures that generated substantial volatility in rankings.

Our results also suggest that it might be desirable for Chile to further explore filtering schemes that would more explicitly exploit the time-series dimension in school performance measures, trying to extract whatever signal they contain. Unfortunately, this is unlikely to be easy and there is not much accumulated experience on how to implement these techniques; they are not in widespread use even in the U.S. Further, it might be worthwhile to invest resources in obtaining the type of data—such as that collected in a handful of U.S. states—that allows calculation of individual-level gains as students progress through the school system. Neither of these strategies is likely to provide a full solution to the challenges described above, but they might help improve the quality and usefulness of rankings.

Finally, we note we cannot be sure to what extent these findings generalize to other settings. If they do to a significant degree, they leave open the possibility that while using information to improve educational service quality along dimensions like absenteeism might be relatively simple, doing so to improve educational quality more broadly may be more of a challenge. This is important because the modal developing country would have to rank schools using significantly less information than Chile, so that in many cases it would be hard to even ascertain to what extent these problems are present. In terms of the desirability of using rankings in educational policy, therefore, one size may not fit all.

References

Banerjee, A., Duflo, E., 2006. Addressing absence. Journal of Economic Perspectives 20 (1), 117–132.

Banerjee, A., Deaton, A., Duflo, E., 2004. Wealth, health, and health services in rural Rajasthan. American Economic Review 94 (2), 326–330.

Chaudhury, N., Hammer, J., Kremer, M., Muralidharan, K., Rogers, F.H., 2006. Missing in action: teacher and health worker absence in developing countries. Journal of Economic Perspectives 20 (1), 91–116.

Chay, K., McEwan, P., Urquiola, M., 2005. The central role of noise in evaluating interventions that use test scores to rank schools. American Economic Review 95 (4), 1237–1258 (September).

Das, J., Zajonc, T., Andrabi, T., Khwaja, A., 2006. Report card methodology: LEAPS project. Mimeo.

Duflo, E., Hanna, R., 2005. Monitoring works: getting teachers to come to school, mimeo, Massachusetts Institute of Technology.

Glewwe, P., Ilias, N., Kremer, M., 2003. Teacher incentives. Working Paper, vol. 9671. National Bureau of Economic Research, Cambridge, MA.

Hanushek, 2004. United states lessons about school accountability. CESifo DICE Report. Winter.

Hanushek, E., Kain, J., Rivkin, S., 2004. Disruption versus Tiebout improvement: the costs and benefits of switching schools. The Journal of Public Economics 88, 1721–1746.

Hsieh, C., Urquiola, M., 2006. The effects of generalized school choice on achievement and stratification: evidence from Chile's school voucher program. The Journal of Public Economics 90 (8–9), 1365–1764.

Kane, T., Staiger, D., 2001. Improving school accountability measures. NBER Working Paper, vol. 8156.

Kane, T., Staiger, D., 2002. The promise and pitfalls of using imprecise school accountability measures. Journal of Economic Perspectives 16 (4), 91–114.

Kremer, M., Chen, D., 2001. An interim report on a teacher attendance incentive program in Kenya. Mimeo, Harvard University.

Kremer, M., Vermeesh, C., 2005. School committee empowerment: Preliminary notes. Mimeo, Harvard University.

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

McEwan, P., Santibañez, L., 2004. Teacher incentives and student achievement: evidence from a large-scale reform. Unpublished manuscript, Wellesley College and RAND.

Mizala, A., Romaguera, P., 2000. School performance and choice: the Chilean experience. Journal of Human Resources 35 (2), 392–417.

Mizala, A., Romaguera, P., 2002. Evaluación del desempeño e incentivos en la educación Chilena, Cuadernos de Economía, Año 39, Nro. 118, 2002.

Mizala, A., Romaguera, P., 2004. School and teacher performance incentives: the Latin American experience. International Journal of Educational Development 24, 739–754.
Pritchett, L., 2004. Towards a new consensus for addressing the global challenge of the lack of education. Copenhagen Consensus Challenge Paper.
Reinikka, R., Svensson, J., 2005a. Local capture: evidence from a central government transfer program in Uganda. Quarterly Journal of Economics 679.
Reinikka, R., Svensson, J., 2005b. The power of information: evidence from a newspaper campaign to reduce capture of public funds, mimeo, http://www.iies.su.se/svenssoj/information2005c.pdf.
Skinner, R., Staresina, N., 2004. State of the states. Education Week 23 (17), 97–99.
The World Bank, 2004. Making Services Work for Poor People. World Development Report, Washington, D.C.
Urquiola, M., Verhoogen, E., 2006. Class size and sorting in market equilibrium, mimeo, Columbia University.