Journal of Urban Economics 65 (2009) 314–322

Contents lists available at ScienceDirect

Journal of Urban Economics

www.elsevier.com/locate/jue

Reduced-class distinctions: Effort, ability, and the education production function ☆

Philip Babcock [a,*], Julian R. Betts [b,c]

[a] UC Santa Barbara, United States
[b] UC San Diego, United States
[c] NBER, United States

ARTICLE INFO

Article history:
Received 5 June 2008
Revised 15 January 2009
Available online 6 February 2009

ABSTRACT

Do smaller classes boost achievement mainly by helping teachers impart specific academic skills to students with low academic achievement? Or do they do so primarily by helping teachers engage poorly behaving students? The analysis uses the grade 3 to 4 transition in San Diego Unified School District as a source of exogenous variation in class size (given a California law funding small classes until grade 3). Grade 1 report cards allow separate identification of low-effort and low-achieving students. Results indicate that elicitation of effort or engagement, rather than the teaching of specific skills, may be the dominant channel by which small classes influence disadvantaged students.

1. Introduction

Findings from the Tennessee STAR experiment raise fundamental questions about the education production function. In the STAR data, mean test scores of students exposed to small classes in kindergarten through third grade exceeded those of students in large classes, and percentile gains appeared largest for disadvantaged students.[1] However, percentile gains appeared to fade by the end of high school. Attitudinal changes in minority students, as captured by a higher probability of taking college entrance exams, appear to have been the major long-run effect of small grade-school classes. The mixed quality of these findings—the differing effects on different kinds of students—motivate an exploration of possible mechanisms by which class size may influence education outcomes.

Two distinct strands of thought—metaphors for education production—inform recent work. In the first, the labor force consists of educators. Their purpose is to communicate to students the knowledge of how to perform specific tasks. Students, then, resemble material inputs. Teachers and administrators are the skilled workers whose labor, combined with books, school buildings, and other factors, add value to the material input. Students function as passive recipients of human capital. Call this the "students-as-material" model of education production. In the second framework, the labor force consists of both educators and students. Teachers resemble managers, and students, the workers they supervise on the factory floor. Here, teachers contribute to education production by eliciting high effort choices from their workers. The managers' primary task is to prevent shirking. They accomplish this by instituting the optimal production technology, monitoring techniques, and incentive structures. Call this the "students-as-labor" model of education production.

Both frameworks capture important aspects of education production. Educators teach specific skills to students and they incentivize students. However, policy choices may depend on which causal mechanism one believes to be dominant. If output on the factory floor is low because workers lack incentives, then programs designed to raise skill levels may be less effective at raising output than a restructuring of their pay schedules or increased monitoring. Similarly, incentives do little for workers who simply lack the skill or knowledge to accomplish tasks demanded of them. Which channel drives the outcomes valued by policy-makers and which framework better explains patterns in the data on class size and education outcomes?

One way to shed light on these questions is to analyze a setting in which the stylized predictions of the two frameworks diverge. If teachers lack time to transfer knowledge to students in large classes, then small class sizes might benefit low-performing students disproportionately, because low performers need more time to learn the subject material. If, instead, teachers are middle managers whose task is to elicit effort from workers, then a different

☆ The authors would like to thank Vince Crawford, Roger Gordon, Gordon Hanson and two anonymous referees for helpful comments. This paper makes use of a database generated by an ongoing project in San Diego Unified School District. This underlying database was made possible by past funding for other projects from the William and Flora Hewlett Foundation, the Public Policy Institute of California, the Bill and Melinda Gates Foundation, the Atlantic Philanthropies and the Girard Foundation.

* Corresponding author at: University of California, Santa Barbara, Economics, Mail Stop 9210, Santa Barbara, CA, USA.
E-mail address: babcock@econ.ucsb.edu (P. Babcock).

[1] See Finn and Achilles (1999), Hanushek (1999), Nye et al. (1999), Krueger (1999), and Krueger and Whitmore (1999) for details on the empirical results summarized in this paragraph.

0094-1190/$ – see front matter
doi:10.1016/j.jue.2009.02.001

outcome might be expected. How do teachers elicit effort, and, more to the point, how might small classes facilitate this? It may be that some students do not connect as strongly to the education setting when they are in large classes. This can be thought of as the "school socialization" effect.[2] Smaller classes may allow teachers to incentivize students whose connection to the educational institution would otherwise have been tenuous. Although school socialization may operate in subtle ways, the cost of incentivizing students—the time and effort of the teacher/manager—could be expressed as a monitoring cost. If students are analogous to workers on fixed-rate as opposed to variable-rate pay, then a factory with a high ratio of managers to workers may improve outcomes because it is easier to monitor workers. Workers prone to shirking would be the agents affected by additional monitoring. In the students-as-material framework, effort elicitation is not the dominant channel through which class-size influences outcomes. The prediction is that low-ability students might benefit most from smaller classes. The students-as-labor framework, in contrast, suggests that students with high disutility of effort—those most prone to shirking—would benefit disproportionately from small classes.[3]

This paper uses a panel dataset containing achievement scores, Grade Point Averages (GPAs), and a rich set of behavior measures for primary school students in the San Diego Unified School District to test the divergent implications of the frameworks. While the results from the Tennessee STAR project show an apparent influence of small class sizes in kindergarten through third grade on minority students' decisions to take college entrance exams, there is no direct evidence that this is because minority students are less socialized to schooling. An obvious alternative would be that minority students arrive at school with less human capital.[4] What has been missing is an empirical design that uses direct measures of behavior and attitude in grade school to identify disengaged or low-effort students, rather than relying on race as a weak (and controversial) proxy for disengagement. This paper attempts to fill that gap. In particular, the behavior measures allow us to treat effort as an observable, and so to group students by effort types, as well as grouping them by achievement.

Investigations of the effect of class-size on student achievement outcomes are common in the literature. We focus on a subtly different question here not only because it may fill a gap in the literature, but because our data may be better suited for this question. We believe class size to be endogenous. While our data offer a plausible source of exogenous variation in class-size between grades (the implementation of California State measure SB 1777 reduced class sizes in lower grades), endogenous class-size variation within grade remains a significant problem. By focusing on differences in the effects of the transition to large classes in higher grades for different subgroups of students, we are able to difference out this source of endogeneity.

Using this approach, we find evidence that larger class sizes disproportionately lower the achievement of students who in grade 1 had relatively low behavior grades. This result stands in contrast to our comparison between high and low ability students (as measured by their academic GPA in grade 1). We found no difference between these two groups when they were moved to large classes. We infer that smaller class size may do more to engage low-effort students than to help low-achieving students.

2. Data and empirical strategy

2.1. Data

The dataset for the analysis consists of a panel of students from 127 elementary schools, grades 1 through 5, in the San Diego Unified School District, for the school years 1998–1999 through 2001–2002. Achievement outcome variables include Stanford 9 math and reading scores, and GPA in core subjects. Teacher evaluations of SDUSD elementary school students include assessments of a broad range of potentially relevant behavioral variables, including "begins work promptly," "follows directions," "classroom behavior," "practices self-discipline." The average of these, the "behavior-GPA," will be interpreted as a measure of student effort.[5] Table 1 shows descriptive statistics for variables that will be used in the analysis. We use Stanford 9 test score *gains* for school years 1999–2000, 2000–2001, and 2001–2002 (as the 1998–1999 school year is earliest for which we have scores).

The Stanford 9 test scores are available in several formats. We use the vertically scaled scores—a format that facilitates the between-grade comparisons required in our analysis. These are psychometrically scaled versions of the raw scores. Item Response Theory is used to weight questions of varying difficulty. Scores that are scaled within a grade have the property that a five-point gain represents the same amount of learning at different points in the test-score distribution. Test scores like the Stanford 9 scores we use that are also vertically scaled have the property that the scores are comparable across grades. Thus, for example, a student with the same score in grades 3 and 4 has not progressed, while two students with gains of five points have improved by the same amount. There is a large statistical literature behind this approach. Psychometrics is not without critics, but the data suggest that within elementary schools average gains are quite constant across most grades, which is what we would expect.

Panel B, column 1 shows descriptive statistics for all students in either grade 3 or 4, the grades that will be the focus of the analysis here. Many of these students entered grades 3 and 4 outside of the years that we will be using, as we focus on two cohorts of grade 4 students who entered grade 4 in either 2000–2001 or 2001–2002. The second column (labeled "potential sample") shows the mean and standard deviation for students who were ever in the district enrolled in one of these cohorts, and thus could potentially have attended SDUSD continuously in grades 1–4 during the years for which we have data. There is attrition and new entry, so we cannot compute test score *gains* for all of these students. Specifically, we can measure gains in grade 3 and grade 4 only for students who attended SDUSD for three straight years (in grades 2–4). This is the sample referenced in column 3. Although the sample size drops by about a third, the means of key variables are very similar to those shown in column 2.

Further, we will identify high types and low types based on performance in grade 1 (as described in Section 2.3). So the difference-in-differences analysis is possible only for students who

[2] Recent work in the theoretical literature formalizes notions of "school socialization." Akerlof and Kranton (2002), for example, hypothesize that schools do not simply produce skills, but "impart an image of ideal students, in terms of characteristics and behavior," and that this affords schools the opportunity to elicit (or discourage) effort. See also Bowles et al. (2001), and Carneiro and Heckman (2003) for theoretical explorations of the "socialization" of students through the creation of incentive-enhancing preferences or non-cognitive human capital, respectively.

[3] We mean here that students "benefit" in the eyes of a social planner who wants to raise skill acquisition outcomes, not necessarily that students' utility rises.

[4] See Coley (2002) for evidence of inequality in human capital observed at the time of students' entry into kindergarten. He finds that cognitive development is positively linked to family income.

[5] We use a simple average of the behavior components to construct behavior-GPA. This specific choice of weighting does not appear to drive the results in Section 4. (Likewise, weightings on the components of academic-GPA do not appear to drive results.) Results are qualitatively similar given different weightings; however coefficients on interaction terms involving behavior GPA are sometimes less precisely estimated when components of behavior-GPA are excluded altogether.

Table 1
Pooled sample: 1999–2000, 2000–2001, 2001–2002 school years.

Panel A. Class size: Grades 1–5

	Mean	Std. dev.	Obs
Grade 1	18.52	2.81	34,767
Grade 2	18.86	2.09	33,395
Grade 3	18.62	2.19	32,995
Grade 4	28.50	4.97	32,127
Grade 5	28.69	5.50	30,419
Total (grades 1–5)	22.46	6.11	163,703

Panel B. Descriptive statistics

	Grades 3–4 All[a] (1)		Grades 3–4 Potential sample[b] (2)		Grades 3–4 Attended 2&3&4[c] (3)		Grades 3–4 Attended 1&2&3&4[d] (4)	
	Mean	Std. Dev.			Mean	Std. dev.	Mean	Std. dev.
Sat9 math	622	41	622	41	628	39	627	39
Sat9 read	628	45	628	45	636	43	635	43
Sat9 math (gain)	–	–	26	29	26	29	25	29
Sat9 read (gain)	–	–	31	25	30	25	30	25
Academic-GPA	2.71	.70	2.70	.70	2.81	.66	2.81	.66
Math (letter grade)	2.18	.91	2.19	.92	2.05	.84	2.04	.83
Read (letter grade)	2.83	1.02	2.83	1.02	2.69	.96	2.70	.96
Behavior-GPA	3.07	.84	3.06	.84	3.14	.81	3.13	.82
Begins promptly	3.12	.89	3.11	.89	3.20	.85	3.19	.85
Class behavior	3.09	.90	3.07	.90	3.14	.87	3.13	.88
Self-discipline	3.00	.96	2.99	.96	3.07	.93	3.05	.94
Follows directions	3.07	.89	3.06	.89	3.15	.85	3.14	.86
White (fraction)	.26	.44	.26	.44	.28	.45	.27	.44
Black (fraction)	.15	.35	.14	.35	.13	.34	.14	.34
Hispanic (fraction)	.41	.49	.41	.49	.37	.48	.38	.49
Asian (fraction)	.18	.38	.18	.38	.21	.41	.20	.40
Other (fraction)	.01	.09	.01	.09	.01	.09	.01	.09
Obs	56,494		37,859		24,514		16,784	

[a] This includes all students from lines 3 and 4 of Panel A, except 8628 for whom there is no SAT9 data.
[b] "Potential" sample includes students from column 1 who could potentially have attended SDUSD continuously in grades 1–4 during the years for which we have data. (These are students who entered grade 4 in either 2000–2001 or 2001–2002.)
[c] These students are the subset of the potential sample who attended SDUSD in grades 2 through 4. Migration of students in and out of SDUSD accounts for the difference in sample size between this column and column 2.
[d] These students attended SDUSD in grades 1 through 4. Migration of students in and out of SDUSD accounts for the difference in sample size between this column and column 3.

attended SDUSD in grades 1–4. Descriptive statistics for this sample are reported in column 4, the main subsample used in the paper. Inferences drawn from the main regressions then relate to this subsample, and an important caveat is that these students may differ from students who left or entered the district. Students who attended SDUSD continuously from grades 1 to 4 do appear to have slightly higher test scores than those who did not. However, their test scores and gains look similar to those who attended in grades 2 through 4.

Several factors complicate any analysis of the influence of class size on education outcomes using non-experimental data. In non-experimental settings, there may be no reason to believe that class size is randomly assigned. Administrators may place slower students in smaller classes, in which case reduced-form regressions of achievement on class size could show higher gains in larger classes. Motivated parents of unobservably advantaged students may pressure administrators to place their students in smaller classes, in which case the bias would go in the opposite direction. These effects occur within schools, but between-school sources of endogeneity also exist. Lazear (2001) posits a model in which class size is a choice variable and the optimal class size rises with the attention span of the students. Areas in which students had longer attention spans would then feature larger classes.

Table 2, a first pass at the data, displays results of regressions of math and reading test-score gains on class size and grade dummies, with and without student fixed effects.[6] When there are no student fixed effects, in columns 1 and 2, the coefficient on class size is positive for math and negative for reading, but is not significant in either case. The coefficients from regressions with student fixed effects, in columns 3 and 4, are significantly negative for math and insignificant, very small, and positive for reading. Student fixed effects will mitigate all sources of endogeneity that do not change over time for the individual student, but do not account for the possibility that the size of a student's class in a given year may be related to *changes* in that student's performance or attitude during the previous year (unobserved by the researcher). We conclude that inferences about the relationship between class size and test score gains drawn from these regressions may be problematic.

2.2. Empirical strategy

The research design for this paper will use the transition from third to fourth grade to proxy for a change in class size. In 1996, the California State Legislature passed and began to implement Senate Bill 1777. The purpose of the reform measure was to reduce class size in early grades from what had been an average of

[6] The sample in Table 2 consists of all students in grades 3 through 5 for whom gains are available.

Table 2
Class size and test score gains: Students in grades 3–5.

	OLS		Student fixed effects	
	Dependent variable:		Dependent variable:	
	(1) Math score gains	(2) Reading score gains	(3) Math score gains	(4) Reading score gains
Class size	.00779	−.0340	−.143**	.0220
	(.0234)	(.0208)	(.0580)	(.0503)
Grade 4	−13.2***	−8.25***	−13.6***	−10.8***
	(.371)	(.326)	(.649)	(.563)
Grade 5	−7.98***	−17.5***	−10.0***	−21.5***
	(.351)	(.310)	(.673)	(.583)
R-squared	.04	.09	.06	.10
Root MSE	26.4	23.4	30.8	26.7
Observations	69,926	69,926	69,926	69,926

Based on students in grades 3–5 during 1999–2000, 2000–2001, and 2001–2002 school years. Some students transferred into SDUSD and their records lack test score data for the year prior to the transfer. (25,615 out of 95,541 observations in grades 3–5 lacked data on gains.) This sample differs from the Table 1, Panel B, column 4 sample that will be used in the main analysis, as these basic motivating regressions use data that go beyond fourth grade and do not require early grade 1 data. Robust standard errors in parentheses.

** Significant at 5%.

*** Significant at 1%.

28 students to a maximum of 20. The legislation funded class-size reduction from kindergarten through third grade only. As a result of this legislation, the fourth grade classes have more students, on average, than third grade classes for all school years in the SDUSD dataset (1998–1999 through 2001–2002). Table 1, Panel A, shows class size by grade in San Diego Unified schools. Average class size leaps by 10—from 19 students to 29 students—between third and fourth grade. Class size could still be endogenous, of course, as the law does not mandate exact class sizes. (For grade 3, the measure sets a maximum size but no minimum, and in grade 4 the measure does not impose size requirements.) Thus, we do not use class size as a regressor. We assume only that the transition from grade 3 to grade 4 captures a source of exogenous variation in class size.

This paper will compare math and reading test score gains in third grade to gains in fourth grade for different subgroups of students. If the students-as-material framework captures the relevant causal channel, one would expect low-ability students to exhibit a steeper drop-off in test score gains between third and fourth grade than high-ability students. If the students-as-labor framework applies, then one would expect low-effort or disengaged students to exhibit a steeper drop-off in test-score gains between third and fourth grade than high-effort students. The empirical strategy focuses on a difference-in-differences. While neither the "high" group nor "low" group in these regressions will be a control group, per se, we assume that any important explanatory factors in the transition from third to fourth grade (other than class size) impact test scores for both groups in the same way. The influence of possible confounding factors will be differenced out.

To fix ideas, suppose the expectation of achievement gains in small classes consists of a grade effect, constant across types, and an effect based on one's type (high or low), constant across grades. Suppose that these effects enter additively, so that the conditional expectation may be written

$$E\left[\Delta Y_i^S \mid g, T\right] = \beta_g + \gamma_T$$

where ΔY_i^S is the achievement gain variable for student i in grade g in a small class (S), β_g is the grade effect and γ_T captures the effect of being a type T, which can be either "high" or "low." If big classes (B) alter test score gains of low types differently than they alter test score gains for high types, we may write

$$E\left[\Delta Y_{ig}^B \mid g, T\right] = E\left[\Delta Y_{ig}^S \mid g, T\right] + \delta_T.$$

Here, δ_T varies with type and is the mean difference in test score gains for big classes (ΔY_i^B) relative to small classes. Because classes are big when $g = 4$ and classes are small when $g = 3$,

$$E\left[\Delta Y_{ig}^B \mid g = 4, T = L\right] = \beta_4 + \gamma_L + \delta_L,$$
$$E\left[\Delta Y_{ig}^B \mid g = 4, T = H\right] = \beta_4 + \gamma_H + \delta_H,$$
$$E\left[\Delta Y_{ig}^S \mid g = 3, T = L\right] = \beta_3 + \gamma_L,$$
$$E\left[\Delta Y_{ig}^S \mid g = 3, T = H\right] = \beta_3 + \gamma_H.$$

Subtraction across types and grades yields

$$\left\{E\left[\Delta Y_{ig}^B \mid g = 4, T = L\right] - E\left[\Delta Y_i^B \mid g = 4, T = H\right]\right\} - \left\{E\left[\Delta Y_{ig}^S \mid g = 3, T = L\right] - E\left[\Delta Y_i^S \mid g = 3, T = H\right]\right\} = \delta_L - \delta_H.$$

Here $\delta_L - \delta_H$ represents the difference in the impact of moving to a large class in grade 4 for low types versus high types. If negative, this difference tells us that low types are disproportionately hurt by the move to larger class sizes in grade 4. We will estimate $\delta_L - \delta_H$ by regressing test score gains on grade dummy, type dummy, and grade-type interaction term.[7]

2.3. Defining types

We restrict the sample to students who attended schools in San Diego Unified in grades 2, 3, and 4, so that for each student there are data on test score gains in grade 3 and grade 4. As will be described below, we further restrict the sample to students who attended schools in SDUSD in grade 1, as well. If low types are defined as students with low test scores in grade 3, a problem of regression to the mean arises. The dependent variable is test score gains, $(Y_{i4} - Y_{i3})$ or $(Y_{i3} - Y_{i2})$. Students with high grade 3 test scores (attributable to randomness) will experience systematically lower gains from grade 3 to grade 4 and systematically higher gains from grade 2 to grade 3. If grade 2 test scores are used to identify high types, a similar problem occurs: Students with high grade 2 test scores will experience systematically lower gains from grade 2 to grade 3. A solution would be to use first grade test scores to identify types. Unfortunately, Stanford 9 achievement tests were not administered in grade 1. Thus, we use the average letter grade awarded to the student in academic subjects in grade 1, which we denote academic-GPA, to define high-ability and low-ability types.[8] We define high ability/low ability as students with GPA above/below the district-wide average in grade 1.

To identify high-effort and low-effort types, we use a first grade "behavior-GPA" that is the average of the 4 measures of behavior shown in Table 1.[9] For each case, we define high types as those whose measured outcome (academic-GPA on the one hand, or behavior-GPA on the other) exceeds the school first-grade average.

Our attempt to test whether small classes help primarily low-ability or low-effort students is meaningful only insofar as these are distinct concepts. The plot of academic-GPA against behavior-GPA in Fig. 1 shows many off-diagonal points. Table 3 shows a cross-tabulation of ability type by effort type for students in the

[7] The exposition here follows Angrist and Krueger (1999).

[8] Academic-GPA is the average of core subjects: reading, written language, oral language, spelling, handwriting, English as a Second Language, math, social studies, science, homework, home reading, book reports. (Letter grades are mapped to the customary numerical values, 4 points for an A, etc.)

[9] We use a simple average of the behavior components to construct behavior-GPA. This specific choice of weighting does not appear to drive the results in Section 3. (Likewise, weightings on the components of academic-GPA do not appear to drive results.) Results are qualitatively similar given different weightings; however coefficients on interaction terms involving behavior-GPA are sometimes less precisely estimated when components of behavior-GPA are excluded altogether.

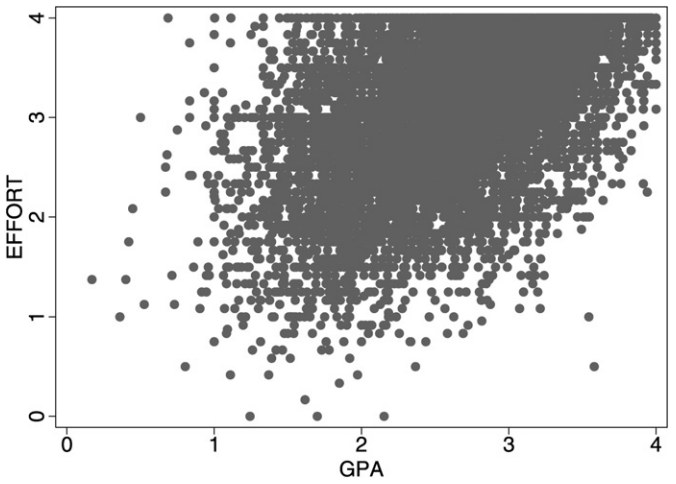

Fig. 1. Academic-GPA vs. behavior-GPA.

Table 3
Effort type–ability type cross-tabulation.

Ability type	Effort type	
	Low	High
Low	4514 (26.9%)	2488 (14.8%)
High	2518 (15.0%)	7264 (43.3%)

Based on Table 1.B, column 4 sample. High (low) effort types are students whose academic-GPA exceeded (did not exceed) the school first-grade academic-GPA average. High (low) ability types are students whose behavior-GPA exceeded (did not exceed) the school first-grade behavior-GPA average.

sample. Thirty percent of the students in the sample lie off the table diagonal, suggesting that effort and ability are related but distinct.[10]

3. Results

3.1. Difference-in-differences estimates by ability, effort, race, and gender

Given that measures of ability, effort, race, and gender are related, we expand the model of Section 2 to include multiple grade-type interactions.[11] Accordingly, regressions in Table 4 include academic-GPA, behavior-GPA, race, gender and grade dummies, along with their associated cross-terms. The coefficients on the grade-type interaction terms estimate differences in differences in test score gains between grades 3 and 4 for high and low types. In row 1, columns 1 and 2, high and low types are based on "ability," as captured by academic-GPA. Low-ability types have academic-GPAs below the average academic-GPA in first grade. Test score gains were smaller in grade 4 than in grade 3, as indicated by the negative coefficient on the grade 4 dummy for columns 1 and 2. In both math and reading, however, the decline in test score gains for low-ability types did not differ significantly from the decline in test score gains for high-ability types.[12]

Row 2 shows difference-in-differences estimates when low types are those who earned below average behavior grades in first grade. This particular metric is meant to capture attitude rather than ability. Estimates for the cross-term in row 2, columns 1 and 2, are negative and significant. Math score gains for low-effort types fell by 2.2 more points between grades 3 and 4 than did the gains of the high effort comparison group. The outcome for reading was similar: Gains of low-effort types fell by 2.15 points more than gains for high-effort types.

How large are these differences in differences? One way to get a sense of the magnitude is to compare the differences in gains by type to district-wide standard deviations in test scores. Grade 4 standard deviations in test scores are 40 and 43 points for math and reading, respectively. So the increase in class size in grade 4 throws low behavior-GPA students about .05 standard deviations further behind high behavior-GPA students in both math and reading. These differences are of the same order of magnitude as effects of small classes that have been reported in previous work.[13]

Breakdowns of type by race and gender do not produce statistically significant differences in differences (rows 3–5). Larger classes, then, appear not to reduce test score gains of low academic-GPA students more than high academic-GPA students, of black students more than non-black students, or of males more than females, but they do appear to reduce gains of low-effort students disproportionately. The results lend support to the students-as-labor framework, as opposed to the students-as-material framework. Small classes may allow teachers/managers to monitor, motivate, and incentivize low-effort students.

One could argue that some other change between grades 3 and 4 drives the observed difference-in-differences for math and reading gains. Perhaps teacher qualifications or experience varies systematically between grades 3 and 4. To account for this possibility we include measures of teacher characteristics as additional covariates in the regressions of Table 4, columns 3 and 4. Dummies in these regressions indicate whether a students' teacher in a given year possessed a master's degree, whether she possessed an emergency certificate, whether she possessed an intern certificate, and whether she had 0–2 years, 3–5 years, or 5–7 years experience teaching. Addition of these controls does not alter results significantly, evidence that systematic differences in teacher characteristics between grades 3 and 4 do not drive the observed difference-in-differences.

Further, if class size is indeed endogenous, it could be that administrators assign high-effort and low-effort students systematically to different-sized classes within grade 4 and within grade 3, and that differences in class size, *within grade*, drive the results above. To account for this possibility, we include class size (in addition to the full set of controls from columns 3 and 4) in the regressions of Table 4, columns 5 and 6. The coefficient on class size is small and insignificant, and low-effort cross-terms do not change in this specification. It would appear, then, that changes in within-grade differences in class size for low-effort and high-effort types do not drive the findings in this section.

[10] "Low effort" and "high effort" might seem to describe choices rather than types. We argue here that effort grades enable us to identify types—students with low and high disutility of effort. We abbreviate, then, when we use the labels "high-effort" and "low-effort" to describe these types. In particular, because significant numbers of students lie off the main diagonal of the correlation table, we argue that it makes sense to think of ability and attitude as separate endowments.

[11] Adding covariates introduces the possibility that correlation between regressors influences the results. Results in this section do not appear to be driven by interaction between ability, effort, race, and gender covariates. Main results persist in regressions based on the simpler model.

[12] It could be that the difference in gains in math scores in large and small class settings varies with abilities specific to math, rather than with our broader measure of ability (academic-GPA). Similarly, the difference-in-differences in reading gains may vary with initial reading ability. Regressions that use these subject-specific measures (math letter grade and reading letter grade), available upon request, produce similar results.

[13] For example, using Project STAR data, Schanzenbach (2007) shows about a .1 standard deviation difference between the test-score benefits of blacks and whites associated with exposure to smaller classes.

Table 4
Difference-in-differences in test score gains by type (with controls).

	Dependent variable:					
	Math score gains (1)	Reading score gains (2)	Math score gains (3)	Reading score gains (4)	Math score gains (5)	Reading score gains (6)
Grade 4 × Low acad-GPA	−.0186	.201	−.0578	.217	−.0593	.216
	(1.13)	(.972)	(1.13)	(.972)	(1.13)	(.972)
Grade 4 × Low behav-GPA	−2.20*	−2.15**	−2.27**	−2.23**	−2.27**	−2.23**
	(1.15)	(.988)	(1.15)	(.987)	(1.15)	(.987)
Grade 4 × Black	1.07	−.899	1.05	−.893	1.04	−.894
	(1.60)	(1.38)	(1.60)	(1.38)	(1.60)	(1.38)
Grade 4 × Hispanic	−1.18	1.37	−1.41	1.25	−1.42	1.24
	(1.13)	(.946)	(1.13)	(.949)	(1.13)	(.950)
Grade 4 × Male	1.31	−1.33	1.27	−1.34	1.26	−1.34
	(1.05)	(.886)	(1.05)	(.885)	(1.05)	(.885)
Grade 4	−15.6***	−9.25***	−15.4***	−9.16***	−15.2***	−9.14***
	(.973)	(.841)	(.975)	(.844)	(1.13)	(.997)
Low acad-GPA	.436	1.81***	.611	1.84***	.610	1.84***
	(.705)	(.620)	(.705)	(.621)	(.705)	(.621)
Low behav-GPA	.599	.374	.733	.434	.731	.434
	(.715)	(.628)	(.713)	(.628)	(.713)	(.628)
Black	−2.7***	−1.90**	−2.27**	−1.75**	−2.27**	−1.75**
	(.996)	(.874)	(.999)	(.876)	(.999)	(.876)
Hispanic	.745	1.56**	1.24*	1.70***	1.24*	1.70***
	(.704)	(.608)	(.712)	(.620)	(.712)	(.620)
Male	−1.85***	.439	−1.86***	.431	−1.86***	.431
	(.653)	(.566)	(.652)	(.566)	(.652)	(.566)
Teach qual controls	No	No	Yes	Yes	Yes	Yes
Class size	No	No	No	No	−.0149	−.00205
					(.0622)	(.0537)
R-squared	.08	.05	.08	.05	.08	.05
Root MSE	26.7	24.0	27.6	23.9	27.6	23.9
Observations	16,784	16,784	16,784	16,784	16,784	16,784

All regressions use Table 1.B, column 4 sample. Teacher quality controls include dummy variables for whether a students' teacher in a given year possessed a master's degree, whether she possessed an emergency certificate, whether she possessed an intern certificate, and whether she had 0–2 years, 3–5 years, or 5–7 years experience teaching. Robust standard errors in parentheses.

* Significant at 10%.
** Significant at 5%.
*** Significant at 1%.

3.2. Behavior gains

Our main goal in this paper is to assess how academic achievement responds to variations in class size, by type of student. However, our findings for low-effort students raise the question of whether class size influences student behavior itself. Table 5 reports difference-in-differences regressions in which gain in behavior-GPA is the dependent variable. These specifications are analogous to those used in the SAT9 test score gains regressions of Table 4. The estimate on the low-behavior-GPA cross-term is negative, as expected, but small and insignificant. There is no strong evidence of disproportionate reductions in effort or behavior by low-behavior-GPA types in large classes. The coefficient on the black-grade-4 interaction term, though small, is positive and significant—apparently suggesting that small classes yield slightly greater behavior gains for non-black students than for black students.

However, there may be a problem of scaling in the behavior gains regressions. Table 6 shows summary statistics for behavior-GPA by grade. It would appear that teachers set behavior norms every year so that the mean behavior-GPA is always about 3 and the standard deviation about .85. This does not pose a problem if behavior-GPA from a single year is used to define high and low types. But if the gain in behavior-GPA is the outcome variable, one might not expect to find differences in gains between grades 3 and 4 by type. The predicted greater dispersion in behavior grades in large classes would not show up because of re-norming.

We cannot rule out the possibility that behavior standards are dependent on age and grade, and were simply not designed for between-grade comparisons. (To our knowledge, the district has never attempted to use behavior evaluations for this purpose.) For this reason, we are more confident drawing inferences from regressions in which test score gains are the dependent variable. We hesitate to draw strong conclusions from the behavior-gains regressions.

The renorming suggested by Table 6 motivates a closer look at behavior-GPA. Might effort measures be subjective in other important ways? Might they be defined by idiosyncratic norms and expectations of teachers, with little connection to any objective standard? If this were the case—that is, if behavior-GPA were teacher-specific—one would expect that the average measured behavior-GPA in schools attended by students from low-SES families would not differ significantly from measured behavior-GPA in high-SES schools. Fig. 2, a scatterplot of average school-level behavior-GPA against the percent of student population eligible for free lunches, shows a strong negative correlation.[14] Low behavior-GPA is observed more prevalently in low-SES schools.[15] This would seem a strong indication that behavior-GPA, though normed by grade, is not a strictly subjective or teacher-specific measure.

[14] The associated regression of average behavior-GPA on "percent free lunch" and constant yields a negative coefficient with t-statistic of −81. (Results in Tables 4, 5, 7, and 8 are not altered significantly by adding covariates for "percent free lunch" and "percent free lunch" interacted with grade.)

[15] This would appear consistent with findings that low-SES schools are more likely to report crimes. (See Barton et al. (1998).)

Table 5
Difference-in-differences in behavior-GPA gains by type (with controls).

	Dependent variable:		
	Behavior-GPA gains (1)	Behavior-GPA gains (2)	Behavior-GPA gains (3)
Grade 4 × Low acad-GPA	−.0371	−.0347	−.0346
	(.0287)	(.0286)	(.0286)
Grade 4 × Low behav-GPA	−.00524	−.00796	−.00809
	(.0296)	(.0295)	(.0295)
Grade 4 × Black	.0887**	.0919**	.0923**
	(.0412)	(.0412)	(.0412)
Grade 4 × Hispanic	.0255	.0237	.0244
	(.0271)	(.0271)	(.0271)
Grade 4 × Male	.0171	.0168	.0169
	(.0252)	(.0251)	(.0251)
Grade 4	−.0155	−.0173	−.0264
	(.0196)	(.0197)	(.0244)
Low acad-GPA	.0263	.0246	.0246
	(.0171)	(.0171)	(.0171)
Low behav-GPA	.0357**	.0369**	.0370**
	(.0177)	(.0177)	(.0177)
Black	−.103***	−.107***	−.106***
	(.0244)	(.0245)	(.0245)
Hispanic	−.0552***	−.0566***	−.0562***
	(.0161)	(.0164)	(.0164)
Male	−.0485***	−.0484***	−.0484***
	(.0150)	(.0150)	(.0150)
Teach qual controls	No	Yes	Yes
Class size	No	No	.000968
			(.00156)
R-squared	.003	.005	.005
Root MSE	.66	.66	.66
Observations	16,784	16,784	16,784

All regressions use Table 1.B, column 4 sample. Teacher quality controls include dummy variables for whether a students' teacher in a given year possessed a master's degree, whether she possessed an emergency certificate, whether she possessed an intern certificate, and whether she had 0–2 years, 3–5 years, or 5–7 years experience teaching. Robust standard errors in parentheses.

** Significant at 5%.

*** Significant at 1%.

Table 6
Summary statistics for behavior-GPA.

	Mean	Std. dev.	Obs
Grade 1	2.97	.837	34,643
Grade 2	3.01	.850	33,295
Grade 3	3.02	.858	32,660
Grade 4	3.00	.881	32,052
Grade 5	3.04	.896	30,123
Total (grades 1–5)	3.01	.864	162,773

Based on Table 1, Panel A sample (1070 out of 163,703 students in the sample were missing data on behavior-GPA).

4. Alternative explanations

4.1. Grade trends

Age or grade trends could explain the difference-in-differences estimates. If the underlying trend in achievement were such that the gap between test-score-gains of low behavior-GPA and high behavior-GPA students widened with each successive grade level, then this fact alone would be enough to explain the results in Table 4. Do differences in test score gains by effort-types widen between grades 4 and 5, despite the fact that there is no significant change in average class size between these grades? A sub-sample of students who attended SDUSD schools in third and fourth grade also attended a school in SDUSD in fifth grade. For these students, it is possible to generate difference-in-differences regressions (analogous to those in Table 4) that focus on the transition

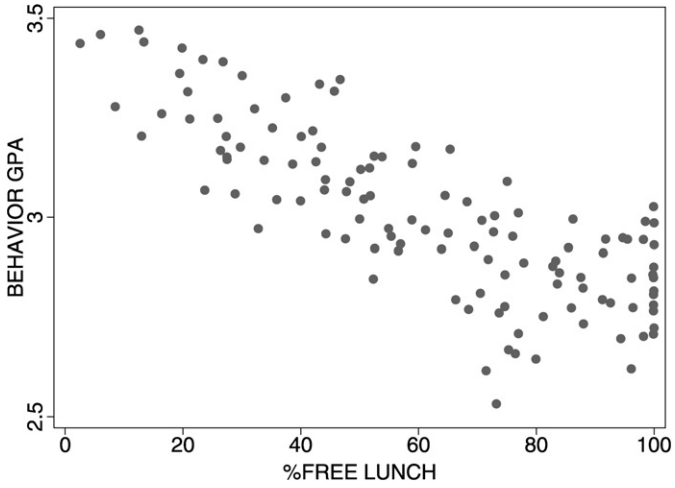

Fig. 2.

from grade 4 to grade 5. These models, then, provide a type of falsification test, and the results are reported in Table 7. The estimates in Table 7 show that the difference in test score gains by behavior-GPA types did not widen significantly either for math or for reading.[16] It would appear that the observed widening of the gap in gains by behavior-GPA types between grades 3 to 4 is not due to an underlying trend that occurs outside of the transition from grades 3 to 4.

4.2. Peer grouping

It could be that the transition from grade 3 to grade 4 leads to systematic changes in peer groupings for low and high types, and if so, could bias our results. Some reflection, and further analysis, suggests that if anything, this confounding factor may be leading us to understate the effect of large class sizes on gains of low-effort students. If there is ability-grouping in classroom assignment, then high types will grouped with high types and low types with low types. When class size rises in grade 4, there are more students in the typical class and fewer classes in the school-grade. When there are fewer classes, administrators group by ability into fewer divisions. As an extreme example, if there were two classes in grade 3 and class sizes rose so that there was only 1 class in grade 4, then ability grouping would disappear altogether in grade 4. Average peer quality would have fallen for high types and risen for low types. If peer effects go in the expected direction, then the merging of classes would increase test score gains for low types relative to high types and decrease the dispersion of outcomes. The observed increase in the dispersion of outcomes for high types and low types, then, obtains in spite of peer effects. The magnitude of the difference-in-differences estimate could be interpreted as a lower bound.

In the data, peer groups evolve as predicted: The evidence suggests that from grade 3 to grade 4 peer quality declined for high-ability types and increased for low-ability types. However, the observed changes are small.[17] Given evidence that in grade 4 the

[16] Further, the finding (of no significant difference-in-differences in test score gains for behavior-GPA types) is robust to specifications that do not include multiple grade-type interaction terms. (See footnote 10.) In Table 7, the difference in reading test score gains by academic-GPA type does appear to widen between grades 4 and 5. But this result appears to be driven by correlation between the grade-type interaction terms, and does not hold in the models that exclude multiple grade-type interaction terms. (Supporting regressions available upon request.)

[17] Specifically, in grade 3, high-ability types (as defined by academic-GPA in first grade) were in classrooms with 71% high-ability types. The percentage of high-ability classroom peers fell to 69% in grade 4 for high-ability types. Between grades

Table 7
Difference-in-differences in test score gains by type using grades 4 and 5 gains as a robustness test for underlying trends.

	Dependent variable:					
	Math score gains (1)	Reading score gains (2)	Math score gains (3)	Reading score gains (4)	Math score gains (5)	Reading score gains (6)
Grade 5 × Low acad-GPA	−1.58	−2.63*	−1.75	−2.6*	−1.71	−2.57*
	(1.58)	(1.38)	(1.57)	(1.38)	(1.57)	(1.38)
Grade 5 × Low behav-GPA	−1.58	2.07	−1.52	2.06	−1.53	2.05
	(1.61)	(1.41)	(1.6)	(1.41)	(1.6)	(1.41)
Grade 5 × Black	1.96	6.13***	2.19	6.15***	2.2	6.16***
	(2.15)	(1.91)	(2.13)	(1.91)	(2.13)	(1.91)
Grade 5 × Hispanic	2.86*	1.74	2.89*	1.71	2.84*	1.67
	(1.55)	(1.3)	(1.55)	(1.3)	(1.55)	(1.3)
Grade 5 × Male	1.75	.854	1.83	.907	1.87	.937
	(1.45)	(1.24)	(1.44)	(1.23)	(1.44)	(1.23)
Grade 5	3.31**	−15.3***	3**	−15.5***	2.74**	−15.7***
	(1.3)	(1.11)	(1.29)	(1.11)	(1.3)	(1.11)
Low acad-GPA	1.95**	3.76***	2.16**	3.89***	2.15**	3.89***
	(.99)	(.83)	(.984)	(.831)	(.984)	(.831)
Low behav-GPA	1.31	−.786	1.39	−.756	1.43	−.727
	(1.01)	(.847)	(1)	(.846)	(1)	(.846)
Black	−2.75**	−2.65**	−2.28*	−2.37**	−2.37*	−2.43**
	(1.34)	(1.17)	(1.32)	(1.17)	(1.32)	(1.17)
Hispanic	−1.62*	2.95***	−1.02	3.28***	−1.16	3.18***
	(.977)	(.782)	(.982)	(.789)	(.984)	(.79)
Male	−1.17	−1.6**	−1.21	−1.68**	−1.26	−1.71**
	(.901)	(.744)	(.896)	(.742)	(.897)	(.742)
Teach qual controls	No	No	Yes	Yes	Yes	Yes
Class size	No	No	No	No	−.174	−.121
					(.0654)	(.0523)
R-squared	.01	.11	.02	.11	.02	.11
Root MSE	25.2	21.4	25.1	23.9	25.1	23.9
Observations	7344	7344	7344	7344	7344	7344

All regressions use sample of students who attended SDUSD continuously in grades 2–5. Teacher quality controls include dummy variables for whether a students' teacher in a given year possessed a master's degree, whether she possessed an emergency certificate, whether she possessed an intern c.

* Significant at 10%.
** Significant at 5%.
*** Significant at 1%.

peer groups of low-ability and low-effort students improved by a small amount, it is worth testing whether these changes in peer groups might have led to an understatement of our main results in earlier tables. Regressions in Table 8 include controls for average first-grade academic-GPA and average first-grade behavior-GPA of classroom peers in third grade (in addition to the full set of covariates in Table 4).[18] As expected, the inclusion of controls for peer quality in the model increases slightly the magnitude of estimated coefficients on the low-effort cross-terms. Point estimates move from −2.27 to −2.29 for math and from −2.23 to −2.26 for reading, suggesting that the absence of controls for peer effects in earlier tables may have biased coefficient estimates toward zero.

Peer effects, then, do not appear to account for the difference in differences in gains by effort-types in small and large classes, but may have caused this difference to be slightly understated.

5. Summary and conclusion

The analysis uses the transition from grade 3 to grade 4 in San Diego Unified as a source of exogenous variation in class size (given a California law funding small classes only up until grade 3). The paper then compares differences in test score gains between grades 3 and 4 for low and high types, using various metrics to define type. Empirical findings indicate that class-size expansion may reduce gains for low-effort students more than for high-effort students, but no significant difference in reductions of gains is observed when types are defined by ability. Underlying grade-level trends do not appear to drive the difference-in-differences findings. Differences in peer grouping for high and low types do not appear to drive the results either, but suggest that the estimated magnitude of the difference-in-differences estimates may be a lower bound. Findings are also robust to the inclusion of controls for possible variation in teacher quality between grades 3 and 4, and for variations in class size in grade 4, itself, among student types.

Previous empirical work on class-size reduction has rarely attempted to look inside the black box and discern the mechanism by which class size may influence education production. In the Tennessee STAR experiment, disadvantaged students appear to have experienced larger test-score gains than advantaged students. A standard explanation is that small classes allow teachers to offer special help to low-achieving students. Results here, if they may be generalized, suggest an alternative explanation—that larger gains for disadvantaged students may have occurred because small classes allow teachers to incentivize disengaged students more effectively, or because students are better able connect to the school setting in small classes.

More generally, findings here suggest it may be important to consider non-cognitive characteristics of students when investigating the effects of increased school resources on student outcomes.

3 and 4, low-ability types saw an increase in the percentage of high-ability types in their classrooms, from 41 to 44%.

[18] It could also be the case that greater *dispersion* of peer quality reduces gains. We have included dispersion measures (standard deviations of the peer measures) in the regressions referenced in Table 8. These did not alter the point estimates or the statistical significance of the main results.

Table 8
Test score gains by type with controls for classroom peers effects.

	Dependent variable:	
	Math score gains (1)	Reading score gains (2)
Grade 4 × Low acad-GPA	−.0116	.211
	(1.13)	(.972)
Grade 4 × Low behav-GPA	−2.29**	−2.26**
	(1.15)	(.987)
Grade 4 × Black	1.05	−.679
	(1.6)	(1.38)
Grade 4 × Hispanic	−1.27	1.36
	(1.13)	(.949)
Grade 4 × Male	1.32	−1.37
	(1.05)	(.886)
Grade 4	−15.4***	−9.17***
	(.976)	(.846)
Low acad-GPA	.279	1.59**
	(.715)	(.631)
Low behav-GPA	.68	.415
	(.715)	(.63)
Black	−2.3**	−1.88**
	(1)	(.878)
Hispanic	1.13	1.59**
	(.712)	(.623)
Male	−1.82***	.499
	(.654)	(.567)
GPA (peer)	−1.12*	−1.2**
	(.602)	(.549)
Effort (peer)	−.575	.158
	(.727)	(.68)
Teacher quality controls	Yes	Yes
R-squared	.09	.05
Root MSE	27.6	23.9
Observations	16,784	16,784

All regressions use Table 1.B, column 4 sample. Teacher quality controls include dummy variables for whether a students' teacher in a given year possessed a master's degree, whether she possessed an emergency certificate, whether she possessed an intern certificate, and whether she had 0–2 years, 3–5 years, or 5–7 years experience teaching. Apart from the addition of covariates acad-GPA (peer) and behav-GPA (peer) and peer dispersion measures, these models are identical to models (3) and (4) from Table 4. Robust standard errors in parentheses.

* Significant at 10%. ** Significant at 5%. *** Significant at 1%.

Not only might some interventions have greater impact on disengaged students (and other groups that policy-makers may wish to target), but the differing impacts of interventions on different types of students may itself provide information about the underlying mechanism. Student attitudes and behaviors may shape the ways in which school spending is transformed into human capital. If so, then empirical research using characterizations of student attitudes and types may be central to the crafting and evaluation of education policy, and to a deeper understanding of human capital production.

References

Akerlof, G., Kranton, R., 2002. Identity and schooling: Some lessons for the economics of education. Journal of Economic Literature 40 (3), 1167–1201.

Angrist, J., Krueger, A., 1999. Empirical strategies in labor economics. In: Ashenfelter, O., Card, D. (Eds.), Handbook of Labor Economics, vol. 3A. Elsevier Science, Amsterdam, pp. 1277–1366.

Barton, P., Coley, R.J., Wenglinsky, H., 1998. Order in the Classroom: Violence, Discipline, and Student Achievement. Educational Testing Service, Princeton, NJ.

Bowles, S., Gintis, H., Osborne, M., 2001. The determinants of earnings: A behavioral approach. Journal of Economic Literature 9 (4), 1137–1176.

Carneiro, P., Heckman, J., 2003. Human capital policy. Working Paper 9495. NBER.

Coley, R.J., 2002. An Uneven Start: Indicators of Inequality in School Readiness. Educational Testing Service, Princeton, NJ.

Finn, J., Achilles, C., 1999. Tennessee's class size study: Findings, implications, misconceptions. Educational Evaluation and Policy Analysis 21 (2), 97–109.

Hanushek, E., 1999. Some findings from an independent investigation of the Tennessee STAR experiment and from other investigations of class size effects. Educational Evaluation and Policy Analysis 21 (2), 143–163.

Krueger, A., 1999. Experimental estimates of education production functions. The Quarterly Journal of Economics 114 (2), 497–532.

Krueger, A., Whitmore, D., 1999. The effect of attending a small class in the early grades on college-test taking and middle school test results: Evidence from project STAR. Mimeo. Princeton.

Lazear, E., 2001. Educational production. Quarterly Journal of Economics 116 (3), 777–803.

Nye, B., Hedges, L., Konstantopoulos, S., 1999. The long-term effects of small classes: A five-year follow-up of the Tennessee class size experiment. Educational Evaluation and Policy Analysis 21 (2), 127–142.

Schanzenbach, D.W., 2007. What have researchers learned from project STAR? Brookings Papers on Education Policy, 205–228.