School Districts and Student Achievement

Matthew M. Chingos, Grover J. Whitehurst, and Michael R. Gallaher

The Brown Center on Education Policy

The Brookings Institution

March 2013

Abstract

School districts are a focus of education reform efforts in the U.S., but there is very little existing research about how important they are to student achievement. We fill this gap in the literature using 10 years of student-level, statewide data on fourth- and fifth-grade students in Florida and North Carolina. A variance decomposition analysis based on hierarchical linear models indicates that districts account for only a small share (1 to 2%) of the total variation in student achievement. But the differences between lower and higher performing districts are large enough to be of practical and policy significance, with a one standard deviation difference in district effectiveness corresponding to about 0.11 standard deviations in student achievement (about 9 weeks of schooling). District performance is generally stable over time, but there are examples of districts that have shown significant increases or decreases in performance.

Introduction

School districts are at the center of public attention and public policy on education reform. Many of the most popular and aggressively promoted school reform efforts are focused at the district level. Performance-based teacher evaluation is a notable case in point. As a condition of competing for funding under the Obama administration's $4.3 billion Race to the Top program, states promised to establish policies requiring school districts to put in place teacher evaluation systems that would heavily weight student achievement gains on state tests. Districts were expected to tie decisions on tenure, promotion, and salary for individual teachers to the resulting evaluation scores. States around the country, including New York, Maryland, Tennessee, New Mexico, and Indiana, are now in the process of requiring districts to implement such teacher evaluation systems, often with short time frames and much of the decisions on design and implementation left to each school district. Presumably individual differences among school districts, certainly including the quality and skills of their management teams, will influence the results.

Many other reform initiatives are focused at the district level in the sense that they are intended to disrupt the school district's monopoly in delivering publicly funded K-12 education services. These include charter schools, vouchers, on-line education, and school portfolio management models. In some sense, these disruptive reforms proceed on the premise that school districts are the irremediable problem rather than the lever for reform, and that the focus on strong leadership from the top misconstrues where the action is on student achievement, which is individual schools, teachers, curriculum, and parental choice of where to educate their children.

The presumed importance of districts is also implied by the media attention given to prominent school superintendents such as Michelle Rhee in Washington, D.C. and Joel Klein in New York City. Rhee, for example, was on the cover of Time Magazine in 2008 with the lead, "Michelle Rhee … head of the D.C. public schools … could transform public education." The pay scale for superintendents also indicates their perceived importance. In New York State, for example, 63 district leaders each received over $300,000 in salary and benefits for the 2011-12 school year, with the superintendent at the top of the list receiving a salary and benefits package of $541,000 (New York State Education Department, 2011).

Private philanthropy has invested heavily in district-level reforms on the premise that districts are a powerful fulcrum for change. One of those philanthropies, the Eli and Edythe Broad Foundation, has led the way in other initiatives that are predicated on the importance of school districts and their leadership. Their annual Broad Prize for Education bestows $1 million on the school district that has shown the best performance and improvement. The Broad Superintendents Academy was founded in 2002 with the goal of finding leaders from both inside and outside education, training them, and having them fill superintendent positions in a third of the 75 largest school districts in the nation. The foundation has not reached that goal, but it has been remarkably successful in placing its graduates in high-level positions: in all, 21 of the nation's 75 largest districts now have superintendents or other highly placed central-office executives who have undergone Broad training (Samuels, 2011).

When we turn from perceptions and intuitions about the importance of school districts to empirical evidence on their impact, we move from a rich to a sparse landscape. Little is known about the impact of school districts on student achievement. And what seems to be known rests on a highly questionable set of methods and assumptions. Among the handful of studies that have addressed the importance of school districts, most have focused on district leadership. The most recent review of the research literature on district leadership, produced by the Mid-Continent Regional Educational Lab (McRel), comes to the conclusion that "district-level leadership matters." This conclusion is based on the authors' finding of "a statistically significant relationship (a positive correlation of 0.24) between district leadership and student achievement" in a meta-analysis of "studies conducted since 1970 that used rigorous, quantitative methods to study the influence of school district leaders on student achievement" (Waters and Marzano, 2007).

This review is a notable example of the misuse of meta-analysis to draw causal conclusions. In particular, nearly all of the 14 studies the authors use for the meta-analysis of the impact of district leaders employed survey methodologies in which samples of superintendents answered questions about their management practices and philosophies. A typical study in the meta-analysis is an unpublished doctoral dissertation in which the means for districts on a state assessment of student academic achievement were regressed on a linear combination of answers provided by superintendents to survey questions. The finding, for example, that a combination of answers to questions about leadership style by district superintendents is associated with differences in student achievement scores is taken as evidence that the leadership style of district leaders is causally related to student outcomes.

By and large the primary studies and the meta-analysis conflate correlation and causation, in particular in that they fail to consider that much of the variation in district performance that is attributed to variables at the district level such as leadership style could be due to differences among districts in the characteristics of the students and families that are served or in the characteristics of the teachers the district employs. The problems with this type of research on district leadership are highlighted in an exchange in the mid-1970s between Bidwell and Kasarda (1975), who conducted such research, and Alexander and Griffin (1976), who criticized it. The former published a model of the role of a variety of district organizational variables on achievement outcomes based on an analysis of data from Colorado school districts. The study is included in the McRel meta-analysis, and has all the causal ambiguities of the other studies in that set.

Alexander and Griffin (1976) published a response to Bidwell and Kasarda (1975) that focused on the omission of control data on levels of student ability at the district level. They control for this input using student IQ test scores in an analysis of data from school districts in Maryland and find that the contribution of district organizational variables of the type studied by Bidwell and Karsada (1975) to aggregate student achievement is substantially reduced or eliminated with this control. Alexander and Griffin (1976) also note that by limiting their analysis to variation between school districts while ignoring variation within districts, within and between schools, and between students, Bidwell and Karsada (1975) are at best accounting for only a fraction of the total variation in student achievement.

The two issues that Alexander and Griffin (1976) identified nearly 40 years ago, failure to control for the socio-economic background of students served by districts and ignoring variation within districts, continues to plague virtually all research on the impact of districts on student achievement. This almost surely leads to false conclusions, including overstating the extent to which student achievement varies across districts by combining variation in actual district effectiveness with variation in the characteristics of the enrolled students. In other words, even if one ignores the problems in causal interpretation that arise from models of the effects of district organization that do not include appropriate controls for student ability, there is still the problem that what the models are accounting for may be such a small portion of the overall variance in student achievement as to be educationally unimportant.

This paper begins to fill this significant gap in the literature by exploring how student outcomes vary across districts using two statewide, student-level longitudinal databases. We do not aim to obtain credible causal estimates of the effect of individual school districts on student achievement. Rather, we explore the associations between school districts and student achievement using the kinds of databases that are now readily available but did not exist when the earlier studies were conducted. We believe that policy decisions made at the state or federal levels having to do with the resources and effort to be placed on district-level reform strategies can be informed by carefully examining the extent to which current variation in student achievement is associated with school districts as organizational units versus schools and teachers.

We use administrative data from the states of Florida and North Carolina to measure how much student achievement varies across observationally similar districts, and put this in the context of variation at the school, classroom, and student levels. We find that district-level variation in Florida and North Carolina accounts for a relatively small fraction of the variation in student achievement, on an order of magnitude of less than 2% of the total variation. But even though district effects are only a small piece of the total variation in student achievement, there are still differences among the academic achievement of demographically similar students in higher and lower performing districts in North Carolina and Florida that are large enough to be of practical and policy significance. A one standard deviation increase in the estimated district effect is associated with an increase in student achievement of 0.10-0.14 standard deviations in math and 0.07-0.11 standard deviations in reading. There are also districts that have displayed noteworthy patterns of performance in terms of student achievement over the last decade, including districts with consistently high or low performance and districts that saw significant growth or declines.

Data

For our analysis we constructed two student-level datasets for Florida and North Carolina and matched students with teachers, schools, and districts for each dataset. Our extract from the Florida Department of Education's K-20 Education Data Warehouse (EDW) contains observations for every student who took state assessments in math and reading from 1998-99 to 2009-10. In addition to student test scores from the Florida Comprehensive Assessment Test (FCAT), the EDW contains information on student demographics, attendance, and program participation—such as the gifted and talented, free and reduced lunch, and English language learner programs. As of 2000-01, students in grade 3-10 took the FCAT in both reading and writing. Therefore, we limit our analysis to the nine years of data between 2000-01 and 2009-10 for both Florida and North Carolina.

In Florida, we used the EDW's course records and matched students with one teacher for math and one teacher for reading, with most students matching to the same teacher for both subjects. Students were only matched to teachers if the student spent 40% or more of their total academic time with that teacher. Additionally, because the focus of our analysis was on student test-score gains, only students that took the FCAT in at least one subject were included in the analysis. Students with duplicate test scores for the same subject in the same semester were excluded from the analysis, though these students made up a small portion of the initial dataset (less than one half of one percent). Of the students who took the FCAT during our period of study, 90% were matched to a math teacher and 92% were matched to a reading teacher.

Similarly, our extract from the North Carolina Education Research Data Center (NCERDC) contains observations for every student who took End of Grade (EOG) assessments in math and reading through 2009-10. Like Florida, student data from North Carolina include test scores on math and reading assessments and student demographics. Unlike Florida, however, we matched students to teachers based on the identity of the EOG assessment proctor for that year, which according to NCERDC accurately identify 95 percent of all classroom teacher assignments (Hoxby and Weingarth 2005).

We limit our analyses to students in grades 4-5, primarily because students in these grades usually have a single classroom teacher whereas older students have multiple subject-specific teachers (and we exclude third-grade students in order to be able to control for prior-year test scores in some models). We could use teachers of a given subject in the analysis of test scores in that subject, but that would ignore any effects of teachers in other subjects (e.g., the effect of the English teacher on math scores, which might partly reflect students' ability to read word problems). Even though districts are the focus of our analysis, the decision of which teachers to associate with students is important because the portion of the variance in outcomes attributable to teachers affects the portion that could be attributable to districts.

Methods

We first measure the variation in student achievement associated with the district, school, and teacher using variance decomposition techniques based on hierarchical linear models (Bryk and Raudenbush, 1988). These models allow us to measure how much student achievement varies at different levels, namely students within classrooms, classrooms within schools, schools within districts, and districts within Florida and North Carolina.

Specifically, we use Stata's `xtmixed` command to estimate a four-level hierarchical linear model of test scores, with students (level 1) nested within classrooms (level 2) nested within schools (level 3) nested within districts (level 4). Following the notation of Raudenbush, Bryk, Cheong, Congdon, and du Toit (2011), the models are:

Level-1 (students): $Y_{ijkl} = \pi_{0jkl} + \sum_{p=1}^{P} \pi_{pjkl}\alpha_{pijkl} + e_{ijkl}$, (1)

where Y_{ijkl} is the test score of student *i* in classroom *j* in school *k* in district *l*, π_{0jkl} is a constant, π_{pjkl} are level-1 coefficients, $\alpha_{pijkl}$is level-1 predictor *p* for student *i* in classroom *j* in school *k* in district *l*, and e_{ijkl} is the level-1 random error, which is assumed to be normally distributed and homoskedastic.

Level-2 (classrooms): $\pi_{pjkl} = \beta_{p0kl} + \sum_{q=1}^{Q_p} \beta_{pqkl}X_{qjkl} + r_{pjkl}$, (2) where $\pi_{pjkl}$are the coefficients from the level-1 model, β_{p0kl} is a constant, β_{pqjkl} are level-2 coefficients, X_{qjkl} are level-2 predictors, and r_{pjkl} are level-2 random effects.

Level-3 (schools): $\beta_{pqjkl} = \gamma_{pq0} + \sum_{s=1}^{S_{pq}} \gamma_{pqsl} W_{skl} + u_{pqkl}$, (3)

where β_{pqjkl} are the coefficients from the level-2 model, γ_{pq0} is a constant, γ_{pqsl} are level-3 coefficients, W_{skl} are level-3 predictors, and u_{pqkl} are level-3 random effects.

Level-4 (districts): $\gamma_{pqsl} = \delta_{pqs0} + \sum_{g=1}^{G_{pqs}} \delta_{pqsg} Z_{gl} + \nu_{pqsl}$, (4)

where γ_{pqsl} are the coefficients from the level-3 model, δ_{pqs0} is a constant, δ_{pqsg} are level-4 coefficients, Z_{gl} are level-4 predictors, and ν_{pqsl} are level-4 random effects.

The simplest implementation of these equations involves no predictors (other than the constants and a dummy variable identifying fifth-grade students) and allows us to estimate the proportion of the variance in reading and math achievement test scores at fourth and fifth grades that is associated with differences between students vs. classrooms vs. schools vs. districts. In this simple case, we calculate the variance at each level as the variance of the random effects, and divide this by the total variance (sum of variances at each level) to obtain the proportion of the variance at that level. We implement the analysis separately by state, subject (math and reading), and school year.

We also estimate versions of these models that include control variables at the student level to account for variation in student characteristics that is correlated with the teacher, school, and district effects. The controls include age, race/ethnicity, cognitive disability status, free and reduced lunch program status, limited English proficiency status, and, for Florida only, whether the parent/student are native English speakers and whether the student was born in the U.S.

These models also include aggregate characteristics of classrooms, schools, and districts to account for the correlation between the student-level covariates and the random effects.[1]

Finally, we estimate models that also control for students' test scores from the prior year. We largely estimate this model because it is the most appropriate model for teacher effects, which we compare to school and district effects. It is not our preferred model for district and school effects because it is likely that districts that have an impact on student achievement do so in all or most grades. If we measure district effectiveness only by the gains generated for students in grades four and five we would penalize districts that impacted student outcomes in earlier grades. We would also net out any district-level variables such as teacher quality that are associated with prior-year test scores. Consequently, although prior-year scores can serve as a proxy for other unmeasured student characteristics, in this case controlling for them will likely cause us to understate the influence of districts on student achievement.

Because our preferred analysis of district effects does not condition on prior-year scores, it potentially captures multiple years of district influence—not the single year typical of most "value-added" type models. In other words, the analysis estimates the extent to which districts serving similar student populations produce better or worse outcomes in fourth and fifth grade, which will reflect impacts both in those grades and persistent impacts from prior grades.

It is important to emphasize that variance decomposition, no matter how cleverly applied, does not lead to point estimates of causal effects. It is the underlying research design, such as random assignment, that permits causal inference. We frequently use terminology that suggests

[1] We selected which aggregate characteristics to include using a method developed by Mundlak (1978). We ran a series of parameter diagnostics to determine which controls were correlated with random effects at each nested level, then determined that the following aggregate characteristics should be included in the final models: district-level free and reduced lunch program status and race; school-level age, race, cognitive disability status, free and reduced lunch program status, whether the parent/student are native English speakers, and whether the student was born in the U.S; and classroom-level age, ethnicity/race, cognitive disability status, free and reduced lunch program status, whether the parent/student are native English speakers, and whether the student was born in the U.S.

causal effects because alternate phrasing would be convoluted. However, our methods are observational and do not support causal conclusions. When we write, for example, about "gaps in student achievement attributable to school districts" we might more accurately write about "associations among student test scores and the districts in which students are educated that remain after accounting for variation in student achievement within districts that is associated with teachers and schools, and with the inclusion of statistical controls for demographic characteristics of students."

HLM is related to but not identical to approaches used by economists to deal with estimation of effects in hierarchical data, specifically fixed effects. We prefer HLM for our present purposes because it is descriptive whereas the econometric models are focused on the causal effect of one level in a multi-level design, e.g., what is the causal effect of teachers on student achievement having fixed the effects of schools? However, as a test of the robustness of the HLM results we also implement fixed effects regressions that calculate average achievement by district adjusted for student characteristics (but not taking into account the nested structure of the data). Fixed effect estimates, like the results of variance decomposition techniques, are not rigorous causal estimates.

We estimate fixed effects models at the district level using the following specification:

$$Y_{ijkl} = \beta_0 + \alpha X_{ijkl} + \nu_l + \epsilon_{ijkl} \, , \qquad (5)$$

where Y_{ijkl} is the test score of student *i* in classroom *j* in school *k* in district *l*, β_0 is a constant, X_{ijkl} is a vector of student characteristics (identical to those used in the variance decomposition analysis) with coefficient vector α, ν_l is a vector of district fixed effects, and ϵ_{ijkl} is a standard zero-mean error term. The coefficients on the district fixed effects are our value-added estimates for the state, test, subject, and year included in the estimation. Below we show that the district fixed effect estimates are highly correlated with the random effect estimates.

Results

The results of the variance decomposition analysis are presented in Tables 1a and 1b for fourth- and fifth-grade students in 2009-10, the most recent year in both the Florida and North Carolina datasets. The first column of Table 1a shows, for a model of math achievement with no control variables other than a grade dummy, the variance of the random effects at each level, as well as the corresponding share of the total variance in Florida. For example, the results for the district level indicate that the district random effects have an estimated variance of 0.015, which is 1.3 percent of the total variance in math achievement. The share of variance explained increases at the lower levels, to 9 percent at the school level, 31 percent at the teacher level, and the balance of 58 percent at the student level or unexplained (i.e. the residual variance, which includes differences across students within classrooms as well as measurement error).

The second column of Table 1a shows results that include controls for student-level demographic covariates, which explain about one-third of the variance in student outcomes.[2] The shares of the variance explained at the district, school, and teacher levels drop to 1, 2, and 11 percent, respectively. This is not surprising given the well-documented correlation between students' demographic characteristics and their districts, schools, and teachers. The relative drop is smaller at the district level than the school level, probably because there is greater within-district sorting of families across schools than across-district sorting given that Florida school districts are coterminous with counties (i.e. geographically large).

[2] We calculate the variance explained by the controls as the difference between the total variance in a null model (with no controls) and the variance explained at the district, school, teacher, and student levels.

In column three we add controls for prior-year scores, which more than doubles the share of variance explained by the controls to 73 percent.[3] It is unsurprising that prior-year scores are the strongest predictor of current-year scores, and that the share of variance explained by districts, schools, and teachers falls once student's prior achievement is taken into account. The share of variance at the district level falls to one-tenth of one percent (i.e. one thousandth of the total variance). But as we discuss above, these estimates likely understate the importance of districts and schools (but not teachers, who usually only instruct a student for a single year).

The analysis of reading scores, which is reported in columns four through six, yields a similar pattern of results. In our preferred model (column 5), the share of variance explained at the district, school, and teacher levels is smaller for reading than for math, which is consistent with prior research showing that formal education has a larger impact on math achievement than on reading achievement because the latter is more influenced by activities in the home. We also obtain similar results from each of the other school years going back to 2001-02, an interesting finding given the various reforms implemented in Florida during this decade and the substantial improvement in overall student achievement that occurred (Chingos 2012).

Table 1b shows corresponding results for the 115 countywide districts in North Carolina. As in Florida, teachers account for more of the variation in student achievement than schools and districts and the three institutional components account for more variance in math than in reading. But districts explain more than twice the share of variance in North Carolina as they do in Florida: 1.9 and 1.3 percent in math and reading, respectively, compared to 0.8 and 0.4 percent in Florida (using our preferred model that controls only for student demographics). This may be due to the fact that North Carolina districts are smaller than Florida districts, on average.

[3] The sample of students changes from columns (2) to (3) due to missing data on prior-year scores. When we run the model in column (2) on the sample of students included in column (3), we obtain qualitatively similar results (not shown).

Consequently, superintendents of smaller districts may more easily be able to change education policies and practices than their counterparts in larger districts. There may also be more idiosyncratic variability in smaller districts, such as the departure of a highly effective principal of a school that accounts for a significant share of enrollment in the district.

A policymaker may be more interested in how important each of the three "institutional" levels is relative to the total variance across just these three levels (ignoring the variation explained by the control variables and the variation across students within classrooms). Table 2 shows the variance shares rescaled in this way, and indicates that the relative importance of the three levels is much less sensitive to model specification than implied by Tables 1a and 1b, especially in Florida. Teachers in Florida consistently account for 75-85 percent of the institutional variance. Once demographics are accounted for, schools explain 12-15 percent of the variance and districts explain the remaining 4-6 percent. In North Carolina, our preferred estimates apportion 20-24 percent of the variance to the district level, 22-27 percent to the school level, and 53-54 percent to the teacher level. In other words, schools appear to be roughly equal in importance to districts in North Carolina (but not in Florida).

Districts explain a relatively small share of the total variation in student achievement, but are there differences among districts in their contribution to student achievement that are large enough to be relevant for policy? Table 3 converts the variances reported in our preferred models in Tables 1a and 1b to standard deviations. These results indicate that a one standard deviation move in the distribution of district effects is associated with an increase in student achievement of 0.07-0.14 standard deviations. Such a difference corresponds to about 7 weeks of schooling in Florida (about one-fifth of a school year) and 10-12 weeks in North Carolina (one-quarter to one-third of a school year).[4]

The finding that districts explain a small share of variance but are potentially important for student achievement may seem counterintuitive at first, but is consistent with the finding that teachers also explain a relatively paltry share of test scores (less than 4 percent in models that control for prior-year scores) but are hugely important for student achievement. For example, a teacher-level variance of 0.039 in math in Florida, about 3.5 percent of the total variance, corresponds to a standard deviation of 0.20 standard deviations (more than 40 percent of a year of learning). Of course this variation corresponds to differences in teacher quality within schools, which averages out to be roughly the same for any given student over time, as compared to the district and school effects which are relative to other districts and schools.

We also compute the best linear unbiased predictions (BLUPs) of the district-level random effects for each state, subject, and year from the hierarchical linear model. The BLUP calculation procedure shrinks the noisier estimates—those based on a level that explains less variance or those based on smaller clusters—toward the mean. The BLUPs and their 95 percent confidence intervals for math scores in 2009-10 are shown in Figures 1 and 2.

There are a number of districts in both states that perform at levels that are above or below the average for districts in the state by a statistically significant margin. In Florida, 10 percent of districts are statistically significantly above average and 7 percent are statistically below average. The comparable numbers in North Carolina are 10 percent above and 14 percent below. This means there are districts that are over- or under-performing on student achievement relative to what might be expected of them given the characteristics of their students.

[4] Standard deviations are converted to weeks of schooling using the average learning gains for fourth and fifth grade reported by Hill et al. (2008), assuming a 180-day (36-week) school year.

The difference in performance between districts is quite large at the extremes: 0.30 student-level standard deviations separate the highest and lowest performing district in the Florida math data whereas the difference is 0.42 standard deviations for North Carolina. Using the same conversion from Table 3, these ranges correspond to 60 percent of a school year in Florida and more than 80 percent of a school year in North Carolina. The random effect estimates for reading (not shown) follow a similar pattern, as we would expect given the high correlation between math and reading performance aggregated to the district level (correlation coefficients of 0.84-0.90 in 2009-10).

These results are robust to replacing the HLM random effects specification with a fixed effects specification. The standard deviations of the district fixed effects estimates for 2009-10 are 0.13 and 0.09 in Florida for math and reading, respectively, and 0.14 and 0.12 in North Carolina. These standard deviations are modestly larger than those from the random effects model because the fixed effects model does not shrink noisier estimates toward the mean (as the random effects model does). The random and fixed effects produce similar rank orderings of districts, with correlation coefficients in the 0.83-0.87 range.

School district performance, as measured by fourth- and fifth-grade reading and math scores, is quite stable over time. Table 4 shows the year-to-year correlations of the estimated random effects (using our preferred model with demographic controls only) for the ten-year period from 2001-01 through 2009-10. The correlation coefficients are around 0.80 in Florida and above 0.90 in North Carolina. In other words, districts with high or low student achievement (conditional on demographics) in a given year tend to also have high or low achievement the following year.

There are some notable exceptions to this pattern in our data. In Whitehurst et al. (2013), we show examples of districts that, instead of having consistently high or low performance, saw steep declines or impressive gains. For example, one North Carolina district that was at or above the 90th percentile of math performance every year from 2000-01 to 2004-05 was below the 10th percentile in every year from 2007-08 through 2009-10. One way to formalize this analysis is to ask how many districts in Florida and North Carolina experienced statistically significant changes (at the 10 percent level) in performance over the decade covered by our data. In other words, how many had 90 percent confidence intervals in 2000-01 that do not overlap with their 2009-10 intervals?

Out of the 67 Florida districts, three experienced significant gains in math over this period and two experienced significant declines. In reading, two Florida districts experienced significant gains and there were the same number of significant declines. Among the 115 North Carolina districts, five experienced gains and six experienced declines in math, whereas four gained and seven declined in reading. These results indicate that significant changes were uncommon, occurring in fewer than 10 percent of districts, but not unheard of.

Discussion

It is unsurprising that student achievement varies more at the levels close to students—teachers and schools—than at the district level. But our results suggest that there are important differences in student achievement across school districts after taking into account differences in student characteristics. Moving from approximately the 30th to the 70th percentile corresponds to 0.07-0.14 standard deviations in fourth- and fifth-grade student test scores, or 20-33 percent of a year of learning.

A few caveats to these results are worth noting. First, it could be the case that each and every Florida district has a sizeable and similar impact on student achievement through a set of very similar practices. If, for example, every district in Florida added about 6 months of academic growth to the students it served, our analysis would not pick this up, depending as it does on examining variation in outcomes across districts. Such a "main effect" for districts could conceivably be wrought by relieving those closer to instructional interactions, such as building principals, from the time they would have to spend on non-instructionally relevant tasks such as providing for student transportation and meals if there were no district administration above them. We note, however, that this model of a school district effect in which the major function of the district is to provide efficient business services, which nearly all do, is very different from the model in which districts compete for great leaders to drive education reform and enhance student achievement.

Second, our model only considers student performance on state math and reading exams in fourth and fifth grades. Districts may well have effects on performance in other grades, other subjects, and on skills not captured by standardized exams. Our analysis will not capture these effects to the extent that they are unrelated to math and reading test scores in elementary school. And as discussed earlier, our observational methodology generates exploratory findings regarding the importance of districts relative to other levels but does not estimate causal effects.

Finally, it should be noted that the results from Florida and North Carolina cannot necessarily be extrapolated to states with numerous smaller districts. Smaller districts may be more likely to vary in their practices given the greater ease with which a superintendent can change the course of a small district containing a handful of schools, as compared to her counterpart in a large district. This theory is supported by a comparison of the results for Florida and North Carolina. North Carolina has more districts than Florida despite being a less populous state, and districts account for more of the variation in student achievement in North Carolina than in Florida. But data from these two states do not allow us to measure whether districts are more important in states with more numerous districts, such as Texas which has over 1,000 districts.

These limitations aside, these results represent the first effort to measure the importance of school districts using student-level databases and modern statistical techniques. As we discuss in Whitehurst et al. (2013), the decade of data from Florida and North Carolina is rife with examples of consistently high and low performing districts, districts that have seen precipitous declines in achievement, and districts that have made transformative progress. Whether certain characteristics of districts, such as policies and leadership, help explain these different patterns is a ripe subject for future research.

References

Alexander, K.L., & Griffin, L.J. (1976). School district effects on academic achievement: A reconsideration. *American Sociological Review,* 41(1), 144-152.

Bidwell, C.E., & Kasarda, J.D. (1975). School district organization and student achievement. *American Sociological Review*, 40(1), 55-70.

Bryk, A.S., & Raudenbush, S.W. (1988). Toward a more appropriate conceptualization of research on school effects: A three-level hierarchical linear model. *American Journal of Education*, 97(1), 65-108.

Chingos, M.M. (2012). The impact of a universal class-size reduction policy: Evidence from Florida's statewide mandate. *Economics of Education Review*, 31(5), 543-562.

Hill, C.J., Bloom, H.W., Black, A.R., & Lipsey, M.W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives*, 2(3), 172-177.

Hoxby, C.M., & Weingarth, G. (2005). Taking race out of the equation: School reassignment and the structure of peer effects. Harvard University, working paper.

Kane, T.J., Rockoff, J.E., & Staiger, D.O. (2007). What does certification tell us about teacher effectiveness? Evidence from New York City. *Economics of Education Review,* 27(6), 615-31.

Mundlak, Y. (1978). On the pooling of time series and cross section data. *Econometrica*, 46(1), 69-85.

New York State Education Department, Administrative Compensation Information for 2011-2012, available at http://www.p12.nysed.gov/mgtserv/admincomp/docs/2011-12_AdminSalDisc_5_11_11_Post_r.xls.

Raudenbush, S.W., Bryk, A.S., Cheong, Y.F., Congdon, R.T., & du Toit, M. (2011). HLM 7: Hierarchical linear and nonlinear modeling. Lincolnwood, IL: Scientific Software International.
Samuels, C.A. (2011). Critics target growing army of Broad leaders. *Education Week*, June 8, 2011.

Samuels, C.A. (2011). Critics target growing army of Broad leaders. *Education Week*, June 8, 2011.

Waters, J.T., & Marzano, R.J. (2007). School district leadership that works: The effect of superintendent leadership on student achievement. *ERS Spectrum*, 25(2), 1-12.

Whitehurst, G.J., Chingos, M.M., & Gallaher, M.R. (2013). Do school districts matter? Washington, DC: Brown Center on Education Policy, Brookings Institution.

Figure 1. Random-Effect Estimations and 95 Percent Confidence Intervals, Math Achievement, Student-Level Standard Deviation Units, Florida, 2009-10

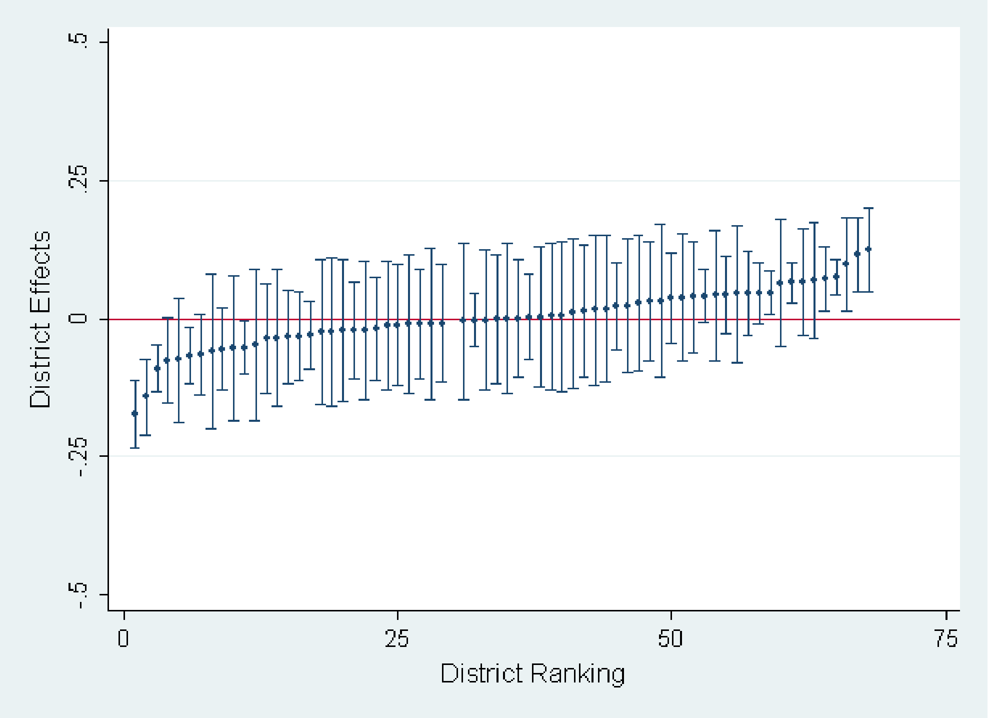

Figure 2. Random-Effect Estimations and 95 Percent Confidence Intervals, Math Achievement, Student-Level Standard Deviation Units, North Carolina, 2009-10

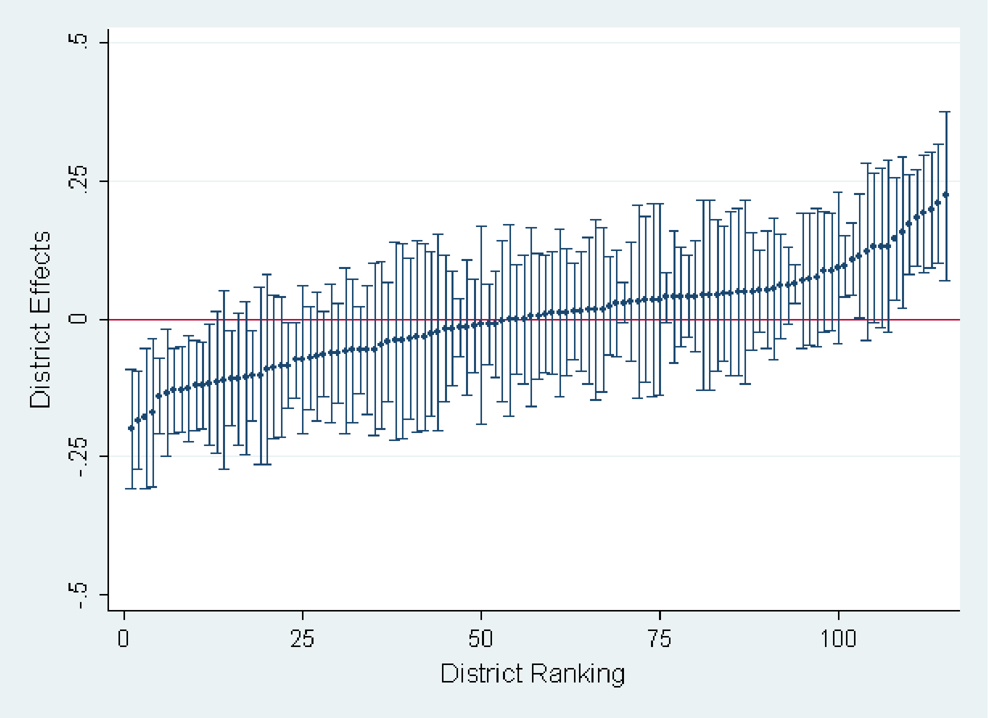

Table 1a. Variance Decomposition of 4th- and 5th-Grade Student Achievement, Florida, 2009-10

	Math			Reading		
	(1)	(2)	(3)	(4)	(5)	(6)
District level	0.015	0.009	0.001	0.017	0.005	0.001
	1.3%	0.8%	0.1%	1.5%	0.4%	0.1%
School level	0.099	0.023	0.007	0.089	0.012	0.003
	8.9%	2.1%	0.7%	8.1%	1.1%	0.3%
Teacher level	0.348	0.120	0.039	0.328	0.089	0.018
	31.4%	10.8%	3.5%	29.6%	8.0%	1.6%
Student level	0.646	0.597	0.255	0.657	0.606	0.306
(residual variance)	58.3%	53.9%	23.0%	59.3%	54.7%	27.6%
Controls		0.359	0.805		0.396	0.780
		32.4%	72.7%		35.8%	70.4%
Demographic controls?	No	Yes	Yes	No	Yes	Yes
Prior-year scores?	No	No	Yes	No	No	Yes
Observations	304,168	304,168	286,677	339,783	339,783	320,554

Notes: Estimates of variance of random effects are calculated using hierarchical linear models. Share of variance is calculated as the variance divided by the total variance from the "null model" which includes no control variables. All models include a dummy variable for grade. Demographic controls include student-level variables for age, race/ethnicity, cognitive disability status, free and reduced lunch program status, limited English proficiency status, whether the parent/student are native English speakers, and whether the student was born in the U.S.; district-level free and reduced lunch program (FLRP) status and race; school-level age, race, cognitive disability, free and reduced lunch program status, whether the parent/student are native English speakers, and whether the student was born in the U.S; and classroom-level age, race, cognitive disability, FLRP status, English language parent/student, and U.S. born. Prior-year scores include test scores in both math and reading, dummy variables for the student's grade in the prior year, and interactions between these dummies and the prior-year scores.

Table 1b. Variance Decomposition of 4th- and 5th-Grade Student Achievement, North Carolina, 2009-10

	Math			Reading		
	(1)	(2)	(3)	(4)	(5)	(6)
District level	0.048	0.020	0.003	0.042	0.013	0.002
	4.8%	1.9%	0.3%	4.2%	1.3%	0.2%
School level	0.098	0.026	0.011	0.094	0.012	0.003
	9.7%	2.6%	1.1%	9.3%	1.2%	0.3%
Teacher level	0.081	0.051	0.032	0.058	0.030	0.011
	8.0%	5.1%	3.2%	5.7%	3.0%	1.1%
Student level	0.782	0.633	0.236	0.815	0.651	0.278
(residual variance)	77.5%	62.7%	23.4%	80.8%	64.6%	27.6%
Controls		0.280	0.727		0.302	0.715
		27.7%	72.0%		29.9%	70.9%
Demographic controls?	No	Yes	Yes	No	Yes	Yes
Prior-year scores?	No	No	Yes	No	No	Yes
Observations	208,485	208,480	194,088	206,649	206,644	193,555

Notes: Estimates of variance of random effects are calculated using hierarchical linear models. Share of variance is calculated as the variance divided by the total variance from the "null model" which includes no control variables. All models include a dummy variable for grade. Demographic controls include student-level variables for age, race/ethnicity, cognitive disability status, free and reduced lunch program status, and limited English proficiency status; district-level free and reduced lunch program (FLRP) status and race; school-level age, race, cognitive disability, and FLRP status; and classroom-level age, race, cognitive disability, and FLRP status. Prior-year scores include test scores in both math and reading, dummy variables for the student's grade in the prior year, and interactions between these dummies and the prior-year scores.

Table 2. Institutional Variance Shares, 4th- and 5th-Grade Student Achievement, 2009-10

	Florida					
	Math			Reading		
	(1)	(2)	(3)	(4)	(5)	(6)
District level	3.1%	6.1%	3.1%	3.9%	4.4%	2.6%
School level	21.5%	15.2%	15.2%	20.6%	11.5%	13.2%
Teacher level	75.4%	78.7%	81.7%	75.5%	84.1%	84.2%
Demographic controls?	No	Yes	Yes	No	Yes	Yes
Prior-year scores?	No	No	Yes	No	No	Yes
Observations	304,168	304,168	286,677	339,783	339,783	320,554

	North Carolina					
	Math			Reading		
	(1)	(2)	(3)	(4)	(5)	(6)
District level	21.2%	20.2%	6.8%	21.7%	23.8%	10.8%
School level	43.3%	27.1%	23.5%	48.4%	21.9%	19.0%
Teacher level	35.5%	52.7%	69.6%	29.9%	54.4%	70.2%
Demographic controls?	No	Yes	Yes	No	Yes	Yes
Prior-year scores?	No	No	Yes	No	No	Yes
Observations	208,485	208,480	194,088	206,649	206,644	193,555

Notes: See notes to Tables 1a and 1b. Percentages are calculated as the variance of the random effects at each level divided by the total variance at the three institutional levels (district, school, and teacher).

Table 3. Distribution of District-Level Random Effect Estimates, 4th- and 5th-Grade Student Achievement, 2009-10

	Florida		North Carolina	
	Math	Reading	Math	Reading
Standard deviation, in test scores	0.10	0.07	0.14	0.11
Standard deviation, in weeks of schooling	7.2	6.8	10.4	11.5

Notes: Standard deviations, calculated in student-level test-score standard deviations, are calculated as square root of variances reported in columns 2 and 5 of Tables 1a and 1b. These statistics are converted to weeks of schooling by dividing by 0.485 in math and 0.36 in reading (the average learning gain between for 4th- and 5th-grade students reported by Hill et al. [2008]) and multiplying by 36 (the number of weeks in a typical 180-day school year).

Table 4. Year-to-Year Correlations of Random and Fixed Effect Estimates, 2000-01 through 2009-10

	Florida		North Carolina	
	Math	Reading	Math	Reading
Unweighted	0.82	0.78	0.92	0.92
Weighted	0.87	0.83	0.92	0.94

Notes: Coefficients indicate linear correlation between best linear unbiased predictions of district-level random effects (for 4th- and 5th-grade student achievement) in adjacent years. Weighted estimates are weighted by district enrollment.