Stanford cepa | Center for Education Policy Analysis

CEPA Working Paper No. 17-02

How much do test scores vary among school districts? New estimates using population data, 2009-2015

AUTHORS

Erin M. Fahle
Stanford University

Sean F. Reardon
Stanford University

ABSTRACT

This paper provides the first population-based evidence on how much standardized test scores vary among public school districts within each state and how segregation explains that variation. Using roughly 300 million standardized test score records in math and ELA for grades 3 through 8 from every U.S. public school district during the 2008-09 to 2014-15 school years, we estimate intraclass correlations (ICCs) as a measure of between-district variation. We characterize the variation in the ICCs across states, as well as the patterns in the ICCs over subjects, grades and cohorts. Further, we investigate the relationship between the ICCs and measures of racial and socioeconomic segregation. We find that between-district variation is greatest, on average, in states with high levels of both white-black and economic segregation.

VERSION

September 2017

Suggested citation: Fahle, E., & Reardon, S.F. (2017). How much do test scores vary among school districts? New estimates using population data, 2009-2015 (CEPA Working Paper No.17-02). Retrieved from Stanford Center for Education Policy Analysis: http://cepa.stanford.edu/wp17-02

How much do test scores vary among school districts?

New estimates using population data, 2009-2015

Erin M. Fahle

Sean F. Reardon

Stanford University

DRAFT: September 2017

Authors Note

Direct correspondence to Erin Fahle (efahle@stanford.edu). This research was supported by grants from the Institute of Education Sciences (R305D110018 and R305B090016), the Spencer Foundation (Award #201500058), and the William T. Grant Foundation (Award #186173) to Stanford University (Sean F. Reardon, Principal Investigator). The paper benefited from collaboration with Andrew Ho, Katherine Castellano, and Ben Shear. We thank Demetra Kalogrides for excellent research assistance. Some of the data used in this paper was provided by the National Center for Education Statistics (NCES). The opinions expressed here are ours and do not represent views of NCES, the Institute of Education Sciences, or the U.S. Department of Education.

Abstract

This paper provides the first population-based evidence on how much standardized test scores vary among public school districts within each state and how segregation explains that variation. Using roughly 300 million standardized test score records in math and ELA for grades 3 through 8 from every U.S. public school district during the 2008-09 to 2014-15 school years, we estimate intraclass correlations (ICCs) as a measure of between-district variation. We characterize the variation in the ICCs across states, as well as the patterns in the ICCs over subjects, grades and cohorts. Further, we investigate the relationship between the ICCs and measures of racial and socioeconomic segregation. We find that between-district variation is greatest, on average, in states with high levels of both white-black and economic segregation.

Keywords: between-district variation, intraclass correlation, segregation

How much do test scores vary among school districts?

New estimates using population data, 2009-2015

Average student academic performance varies substantially among school districts in the United States. This is obvious from even the most cursory glance at publicly available test score data. But what factors shape this variation? To what extent is this variation due to differences in school quality and to what extent is it due to the myriad of out-of-school factors that shape children's opportunities, including their family resources, neighborhood conditions, preschool programs, and afterschool activities?

Answering these questions requires, first, a comprehensive description of the degree and patterns of variation in academic performance among school districts in the U.S. We do not currently have such a description, however. The National Assessment of Educational Progress (NAEP) tests provide information on variation in academic performance among states, but not among schools or districts. State accountability tests can provide school or district-level information, but a comprehensive analysis is complicated by the fact that most states use different standardized tests and that publicly available data often do not include detailed information on each school or district's test score distribution. Other nationally representative studies conducted periodically by the National Center for Education Statistics (NCES) provide estimates of the variation in test scores among schools (e.g., Hedges & Hedberg, 2007), but cannot—because of their sampling designs—describe variation among districts, within individual states, or across a range of grades and student cohorts.

In this paper, we provide a comprehensive description of the patterns of between-district test score variation in the U.S. We apply newly developed methods to estimate the proportion of total test score variance that lies between districts in each state, using roughly 300 million standardized math and English Language Arts (ELA) student test scores from every public school district in the U.S. during the 2008-09 through 2014-15 school years. As we demonstrate below, there is considerable variation among states in the degree to which test scores vary between school districts, and most of this variation is associated with factors outside of school districts' control: between-district variation is greatest, on average, in states with high levels of between-district racial and economic segregation.

Background

Why might test scores vary among school districts? At the district (rather than individual) level, students' average test scores are a function of their accumulated educational opportunities in and out of school to learn the tested material. To the extent that these opportunities vary systematically among school districts, district-average test scores will reflect this variation.

Educational opportunity is closely tied to students' socioeconomic status, and differences among districts in the socioeconomic composition of their student populations is likely a key source of between-district variation in test scores. A large body of literature shows that family socioeconomic status is a strong predictor of academic performance – students from poorer families perform less well on tests relative to their affluent peers (e.g., Reardon, 2011; Sirin, 2005). Even absent any contextual effects on academic achievement, the student-level correlation between family socioeconomic status and achievement will yield a corresponding district-level correlation. And greater economic segregation – more variation among districts in socioeconomic background – will exacerbate the variation among districts in average test scores.

There are many other factors that might similarly shape differences in average achievement across districts through economic or racial segregation – some of which are correlated with local median family income. These include differences in neighborhood conditions and resources (e.g., Sampson, Sharkey, & Raudenbush, 2008; Wodtke, Elwert, & Harding, 2016), differences in the availability and quality of child care, preschool, and afterschool programs (e.g., Chetty et al., 2011; Magnuson, Meyers, Ruhm, & Waldfogel, 2004), and—of course—differences in the resources, curricula, instructional practices, and other dimensions of the quality of local schools (e.g., Card & Krueger, 1992; Deming, Hastings, Kane, & Staiger, 2014). Each of these family, school, and neighborhood factors may independently affect academic performance, but their effects may interact as well. For example, low-income communities may have lower average achievement than high-income communities not only because poor families can afford to provide their children with fewer educational resources at home, but also because concentrated poverty may lead to lower-quality preschool options (Valentino, forthcoming) and lower-quality public K-12 schools. High-income communities, where parents can afford to pay for high-quality childcare and preschool, may also be able to attract more skilled K-12 teachers (Lankford, Loeb, & Wykoff, 2002).

There is clear evidence that test scores vary among *schools* from national studies. For example, using data from the Early Childhood Longitudinal Study, the Longitudinal Study of American Youth, and the National Educational Longitudinal Study, Hedges & Hedberg (2007) find unconditional between-school intraclass correlations (ICCs) of approximately 0.17 to 0.27, depending on the grade-level and subject. In other words, they find that one-fifth to one-quarter of the total variance in test scores on these national assessments is between-schools. Similarly, using the National Assessment of Education Progress data, Konstantopoulos (2009) finds ICCs ranging from 0.10 to 0.25. There have also been a subset of studies looking at between-school variance in particular states. Westine, Spybrook, and Taylor (2013) estimate ICCs using student-level data from Texas in grades 5, 8, 10 and 11. They find that ICCs range from .10 to .20 depending on the subject (science, reading, or math) and grade. Using data from 11 states, Hedges & Hedberg (2014) show that between-school ICCs vary among states from 0.05 to 0.20 across subjects and grades. And, notably, for some states (e.g., West Virginia) they find that these estimates differ significantly from the national between-school ICC estimates in their prior work (Hedges & Hedberg, 2007).[1]

There is little work, however, exploring how much between-*district* test score variation exists and what factors explain this variation. Only one paper to our knowledge uses population data from a set of states to estimate between-district variation. Hedges and Hedberg (2014) estimate both between-school and between-district ICCs for multiple grades in both math and ELA for 11 states each in a single year. The authors show that between-district ICCs are generally much smaller than between-school ICCs – more of the total variation in test scores is between schools than between districts within a state. For example, in Grade 3 math, they find an average between-district ICC of 0.049, compared to a between-school ICC of 0.112. However, their study demonstrates that there is significant variation among states in between-district ICCs. In some states, average student performance was quite similar across districts; in others, performance varied considerably among districts. Hedges and Hedberg find two further patterns in the ICCs: that ICCs are generally larger in math than in reading; and, that between-district ICCs are, on average, larger in later grades than earlier ones. The authors, however, do not systematically explore factors that explain these patterns of variation, and would be limited in their ability to do so given their sample of only a single year of data from each of only 11 states.

Research Questions and Hypotheses

In this paper, we seek to answer two central research questions: (1) How systematically does the amount of between-district variation in test scores vary across states, subjects, grades, and time? and, (2) To what extent is this variation explained by the structure of school districts and the amount of racial and economic segregation between districts?

In line with prior work, we pose two hypotheses about why a relationship between segregation and between-district test score variance would exist. First, as prior work shows, family background and neighborhood conditions exert a strong influence on academic performance—particularly on the development of academic skills in early childhood and elementary school. This would lead to higher average performance in affluent school districts compared to poorer ones, and therefore to more test score variation between districts when the between-district socioeconomic dispersion is wider. Second, school quality may be correlated with local socioeconomic and racial composition. If the schools in poor and predominantly black and Hispanic school districts are inferior, on average, to those in affluent and predominantly white districts, this would also lead to the correlation between segregation and between-district test score variation. Together, these suggest that between-district variation will be higher in more economically and racially segregated states – states with larger disparities among districts in total resources both in and out and school.

We further hypothesize that the effects of exposure to differential educational resources may compound over time. If so, between-district test score variation should be larger in later grades relative to earlier grades. The logic is that local resources (whether in the home, the neighborhood, or the schools) affect academic achievement growth. As a result, not only do resource-poor students start school behind their higher-resource peers, but their academic performance grows more slowly. Therefore, differences in academic achievement related to differences in resources will be smaller when students are younger and grow as they age. This hypothesis suggests that. If true, between-district disparities in test performance should widen faster, on average, as children progress through school in more segregated states than in less segregated states.

Data and Measures

Test Score Data

The test score data in this study come from the federal ED*Facts* data collection system. The data were provided to us by the National Center for Education Statistics under a restricted data use license. The ED*Facts* data include counts of students in each of several ordered proficiency categories (labeled, for example, as “below basic,” “basic,” “proficient,” and “advanced”), by school, year, grade, and test subject for all fifty states and the District of Columbia. Complete data, including math and ELA scores, are available for all tested students in third through eighth grade from the 2008-09 school year through 2014-15, with a small number of exceptions due to non-report, pilot testing, or other extenuating circumstances (e.g. hacking). The full dataset represents students’ scores on roughly 300 million standardized tests administered during this seven-year period.

We aggregate the data to produce counts of students in each proficiency category within each school district-year-grade-subject cell. For each grade, we define a school district as the set of public (including both charter and non-charter) schools that serve students in that grade which are located within the geographic boundaries of a traditional (non-charter) public school district. Operationally, this means we assign charter schools to the traditional (non-charter) local education authority (LEA) in which they are geographically located. By this definition, a district's test score distribution describes the distribution of academic performance of all public school students attending school in a geographically-defined community. The average student test score in a district can therefore be thought of as the result of the total set of educational opportunities and constraints available to students in the community from birth through middle school—including opportunities in their homes, neighborhoods, child care and preschool programs, as well as in their local public schools.[2]

We exclude a small subset of the data. First, we exclude Hawai'i and the District of Columbia in all grades and years because each has only a single school district, making the estimation of between-district variation irrelevant. Second, we exclude schools administered by the Bureau of Indian Education (BIE) due to data comparability issues. Third, in some cases not all students in a state took the same grade-level subject test in a given year. In such cases, between-district variation in test scores will be conflated with between-district differences in the proportions of students taking each test. Based on this requirement, we exclude all data from Nebraska in the 2008-09 school year and math data from Nebraska in the 2009-10 school year, as districts were allowed to select their own assessments in these years and subjects. Additionally, we exclude math data for 7th and 8th grades from California and Virginia in all years, and from Texas in the 2011-12 through 2014-15 school years, as students take end-of-course math assessments. Finally, we exclude data with known reporting issues or data from states where less than 95% of enrolled students, as documented by NCES, were tested.

Measure of Between-District Variation in Test Scores

There are two approaches to measuring between-district variation in test scores. One approach estimates the variance of the means of districts' test score distributions relative to the population variance. This variance of means (denoted τ) and the average within-district variance (denoted σ^2) are typically estimated via maximum likelihood; the ICC is then defined as $\tau/(\tau+\sigma^2)$; if test scores are standardized to have a total variance of 1 (i.e., so that $\tau+\sigma^2=1$), then the ICC is simply τ (or, equivalently, $1-\sigma^2$). This is the approach used by Hedges and Hedberg (2014). The ICC defined this way is useful in designing studies that sample participants from multiple school districts, because the sampling variance of parameter estimates (and therefore the statistical power of a study) depend on this ICC (e.g., Hedges & Hedberg, 2007; Jacob, Zhu, & Bloom, 2010; Raudenbush, Martinez, & Spybrook, 2007; Schochet, 2008).

The other approach is a simple analysis-of-variance decomposition that partitions the total variance of test scores into between- and within-district components. The ICC defined this way describes the proportion of test score variance that lies between, rather than within, districts. If we knew the variance (σ_d^2) of test scores in each school district d in a metric in which test scores are standardized within each state-grade-year-subject, the analysis of variance ICC is defined as $1-\sum_d p_d\sigma_d^2$, where p_d is the proportion of students in a state-grade-year-subject who are in district d. This parameter is useful for describing the patterns of academic performance in the population, since it takes into account the size of each school district.

While both definitions of the ICC describe between-district variation in test scores, they are not identical. First, the analysis of variance approach weights districts by their enrollment, while the variance of means approach does not. Second, the variance of means approach is typically estimated by assuming a common within-district variance; the analysis of variance approach requires no such assumption. If all districts are the same size and the within-district variance is the same everywhere, the two approaches estimate the same parameter. When districts are of different sizes, the two approaches do not estimate the same ICC, because the unweighted variance of district means (τ) is not generally equal to the proportion of test score variance that lies between-districts.

We use the analysis-of-variance approach because it directly estimates the parameter we are interested in: the proportion of test score variance in a state that lies between districts. The variance in district means identifies this parameter only if all districts are the same size or have the same within-district test score variance. In most states school districts vary widely in size; the analysis of variance approach implicitly gives little weight to small districts and more weight to the larger districts. Moreover, the analysis of variance approach does not require the unrealistic assumption that all districts have equal test score variance; our methods (described below) allow for the estimation of unique within-district variances.

We use the proficiency category counts in the ED*Facts* data to construct estimates of the between-district proportion of test score variance in each state-grade-year-subject. In order to estimate σ_d^2 from the raw ED*Facts* proficiency data, we use a new adaptation of the heteroskedastic ordered probit model described by Reardon, Shear, Castellano, and Ho (2016). Using this model, we estimate σ_d^2 in each school district and then calculate the ICC using Equation (12) from that paper, shown below.

$$\widehat{ICC} = 1 - \frac{1}{1 + 2\widehat{\overline{\omega_g^2}}} \sum_d p_d \hat{\sigma}_d^2$$

where $\widehat{\overline{\omega_g^2}}$ is the estimated average sampling variance of the natural log of the standard deviations. The term $1 + 2\widehat{\overline{\omega_g^2}}$ in the denominator corrects the estimated ICC for the fact that the within-district standard deviations are estimated with error; without this correction, the estimated ICCs would be too small.

Using both simulations and analyses of real test score data, Reardon et al. (2016) demonstrate that this approach provides nearly unbiased estimates of district-specific test score distributions and between-district ICCs under a wide range of conditions. Although the Reardon et al. ICC estimator is slightly positively biased, they show that the bias is generally very small—less than 0.005—unless all groups are very small (fewer than 100 students per grade), a condition not present in any state when using school districts as the target groups.[3]

We perform all estimation using the -hetop- ("heteroskedastic ordered probit") command (Shear & Reardon, 2016) n *Stata* (StataCorp, 2013). We estimate the between-district ICC in each state-grade-year-subject using the partially heteroskedastic ordered probit model described by Reardon et al. (2016). This model estimates a common variance for all districts with fewer than 50 students per grade, but allows the variances to vary among larger districts. The estimated ICC from this model has a smaller sampling variance and mean squared error than that from a fully heteroskedastic model. In a few states where only two proficiency categories are reported, we fit homoskedastic ordered probit models (constraining the variances in all districts to be the same), since the heteroskedastic model requires data with at least three ordered proficiency categories.[4]

In total, we estimate 3,795 between-district ICCs from the 49 states in our sample. On average, we produce 77 ICC estimates per state of a maximum possible 84 estimates = (2 subjects x 6 grades x 7 years). Appendix Table 1 shows the number of grade-year-subject estimates in our data by state. To all estimates, we apply a standard measurement error correction of $\frac{1}{r}$, where r is the reported test reliability for the test used in that state-grade-year-subject.[5]

State-level covariates

In states with many small school districts, Tiebout sorting processes (Bayer et al., 2004, 2007; Tiebout, 1956) might lead to low within-district variance in test scores relative to states where most students are concentrated in a few large school districts. We therefore include in our regression models the number of school districts, the average district enrollment, and the Hirschman-Herfindahl Index (HHI) (Herfindahl, 1950; Hirschman, 1964; Hirschman, 1945) of school district enrollments. The HHI measures the extent to which students are concentrated in few large districts or many small ones. In the education literature this is often referred to as a measure of school district fragmentation (Bischoff, 2008; Owens, 2016).[6] We compute these three statistics using data from the Common Core of Data (CCD)[7] separately for each grade (3 through 8) in every year (2008-09 through 2014-15). We then average over grades and years within states to construct four state-level measures.[8]

To measure segregation among school districts, we compute the between-district white-black, white-Hispanic, and poor-non poor (using free lunch receipt as an indicator of poverty) information theory index (H) (Massey & Denton, 1988; Theil & Finezza, 1971) using CCD data.[9] Again, we compute the segregation measures separately by grade and year, and then average each within states.

In our regression models we use the natural logarithm of the number of districts and mean district enrollment. Additionally, we use a transformation of the fragmentation measure: $\ln\left(\frac{1}{1-HHI}\right)$. All variables are transformed after averaging over grades and years. These transformations improve model fit by linearizing the associations between each of the structural covariates and the ICC. Nonetheless, our substantive results are unchanged if we use the untransformed measures (results not shown).

[Table 1]

Summary statistics for all of the state-average covariates and transformed state-average covariates used on the models to improve fit are shown in Table 1. There is significant variation in the structure of school districts across states. Specifically, states range from having approximately 15 school districts (Delaware) to over 1,000 school districts (Texas), with an average of approximately 265 school districts. Correspondingly, the mean grade-level enrollment and standard deviation of grade-level enrollment vary quite significantly across states with some states having all small districts, others having all large districts, and the rest having a mix of both. The HHI ranges from approximately 0.43 (in Nevada) to 0.99 (in a number of states); however, most states have a value above 0.90, reflecting that for almost all states the probability that two randomly selected students are enrolled in different school districts is very high. For the segregation measures, the ranges indicate that in some states there is very little between-district racial and economic segregation (minimum values of each statistic ≤ 0.06), whereas in others there is quite dramatic between-district white-black segregation (maximum value = 0.52), white-Hispanic segregation (maximum value = 0.45) and economic segregation (maximum values = 0.31). Values near 0.5 indicate that on average, each district has only half the diversity of the population as a whole, whereas values less than 0.05 indicate that on average, districts are at least 95% as diverse as the population as a whole. Generally, states with more racial segregation have more economic segregation (pairwise correlations of 0.72 – 0.85) and states with more white-black segregation have more white-Hispanic segregation (correlation of 0.72). A more detailed table of these covariates by state can be found in our online appendix (Appendix Table 1).

Models

The data consist of 3,795 estimated ICCs, nested in 49 states and varying across grades, years, and test subjects. To accommodate the nested data structure, and to take into account the varying sampling variance in the estimated ICCs, we fit precision-weighted random coefficients models to estimate the parameters of interest. The models include an intercept (denoted γ_{00}), a vector of 11 cohort dummy variables each corresponding to the year a cohort of students was in fall of Kindergarten (denoted $\mathbf{C}$; we include dummies for the 2001 to 2011 cohorts and omit the dummy for the 2000 cohort) and a continuous variable denoting grade, and a set of state random effects that allow the ICCs and the linear component of their grade and cohort trends to vary among states. Specifically, the models have this form (one model for each test subject):

$$\widehat{ICC}_{sgy}^{subject} = \gamma_{00} + \mathbf{C} + u_{0s} + (\gamma_{10} + u_{1s})g_{sgy}^{*} + (u_{2s})c_{sgy}^{*} + e_{sgy} + r_{sgy}$$

$$r_{sgy} \sim N\left[0, \hat{v}_{sgy}\right];\ e_{sgy} \sim N[0, \sigma^2];\ \begin{bmatrix} u_{0s} \\ u_{1s} \\ u_{2s} \end{bmatrix} = \boldsymbol{u_s} \sim N[\mathbf{0}, \boldsymbol{\tau}] \tag{2}$$

where $\widehat{ICC}_{sgy}^{subject}$ is the ICC estimate for a state-grade-year case in a given subject; g_{sgy}^{*} is the grade (centered at 5.5); and c_{sgy}^{*} is the student cohort (centered at 2005.5). Note that we can include both a complete set of cohort dummy variables in the model and a random linear cohort term because the linear cohort term has no fixed component; the cohort dummy variables allow the average pattern across cohort to be estimated non-parametrically, while the random linear cohort term allows this nonparametric trend to differ by a linear component among states.

We assume that the estimation error r_{sgy} is normally distributed with zero mean and known variance equal to $\hat{v}_{sgy}$, the estimated sampling variance of $\widehat{ICC}_{sgy}$; the within-state residual error e_{sgy} is normally distributed with mean zero and variance σ^2 to be estimated; and the state-level errors u_{0s}, u_{1s}, and u_{2s} have a multivariate normal distribution with zero means and covariance matrix $\boldsymbol{\tau} = \begin{bmatrix} \tau_{11} & \tau_{12} & \tau_{13} \\ \tau_{21} & \tau_{22} & \tau_{23} \\ \tau_{31} & \tau_{32} & \tau_{33} \end{bmatrix}$, where $\tau_{ij} = \tau_{ji}$ for all $j, i \in [1,3]$. We include the random linear terms on g^* and c^* because we reject the null hypotheses ($p < 0.001$) that the grade and cohort trends do not vary among states (that is, that $\tau_{22} = 0$ and $\tau_{33} = 0$).

Equation 1 describes our baseline model (Model 1), which includes no state-level covariates. In additional models, we add covariates as predictors of the intercept in Model 1 to assess their association with the ICCs. Model 2 includes the structural variables describing the size and number of school districts in a state (transformations of the number of districts, mean grade-level enrollment, and district fragmentation). Model 3 includes the three segregation measures: white-black segregation, white-Hispanic segregation, and free lunch segregation. Model 4 include both the structural variables and segregation measures.

Our final model assesses whether these two sets of covariates are associated with growth in ICCs from third through eighth grade. Using Model 4 as our baseline, we add the structural covariates and the segregation measures as predictors of the grade slope (Model 5). The coefficients on the interactions of the segregation variables and the grade variable indicate the association of segregation with changes in ICCs as cohorts progress through school.

Results

On average, between-district ICCs vary significantly across the U.S. Figure 1 maps the estimated average between-district ICC in each subject. These are the Empirical Bayes estimates from Model 1 (shown in Table 2), though because the reliability of the estimates is over 0.99, there is virtually no shrinkage in these estimates. The ICCs range from near zero (0.009 in ELA, 0.013 in math) to 0.232 in ELA and 0.237 in math. An ICC of 0 implies that all test score variation is within districts (all districts have the same average test score); whereas an ICC of 0.2 means that one fifth of the total within-state variance in test scores is due to between-district differences. This is a relatively large ICC. In such a case, the population-weighted average between-district variance is one-quarter the population-weighted within-district variance (put differently, the district means have a population-weighted standard deviation that is half as large as the average within-district standard deviation of scores).

[Figure 1]

The ICCs are generally larger in math than in ELA by approximately 12% – as shown by the darker coloring in the math map, and intercepts in Table 2. [10] However, despite this difference in magnitude, the correlation between the math and ELA ICCs is 0.943, which means that states with higher between-district variability in math also have higher between-district variability in ELA. This high correlation suggests that the factors that generate more between-district variability within a state are not-subject specific.

[Table 2]

Model 1 provides clear evidence that the ICC increases over grades. The positive growth over grades in both subjects suggests that the factors leading to between-district variation in test scores compound over time. Moreover, the rate of increase varies across states. Some states exhibit negative subject-specific growth rates over grades and others large positive grade slopes. The average increase per grade is approximately 0.0063 in math and 0.0038 in ELA, so from third to eighth grade the average ICC increases by about 0.032 in math and 0.019 in ELA. This is approximately one third of the size of average ICC in math for the 2000 cohort (0.032 / 0.0992 = 0.32 or 32%) and one fifth of the average ICC in ELA for the 2000 cohort (0.019 / 0.0882 = 0.22 or 22%).

For both math and ELA, Models 1 also indicates that ICCs have increased among recent cohorts. Between the 2000 and 2011 cohorts, the average state's ICC increased substantially—by 0.030 (or 31%) in math and 0.031 (or 35%) in ELA. Over the 12 cohorts included in our sample, the trend is nearly linear; in the interest of parsimony, we do not report all of the coefficients on the cohort dummies in Table 2 (though they are available in Appendix Table II).

In both subjects, Model 2 shows that structural differences in district size and enrollment across states explain approximately one third of the variation in the ICCs across states (27-36% depending on the subject). The three structural measures are not, however, jointly statistically significant in Model 2 (p=0.16 in math; p=0.047 in ELA). Model 3 shows that the three segregation measures are jointly statistically significant predictors of both math and ELA ICCs, and together explain 84-86% of the between-state variance (*p*<.01 in both math and ELA). Notably, adding structural covariates to the model (Model 4) explains only slightly more than the segregation measures alone (87-88%), suggesting that segregation is the key factor in explaining variance in the ICCs among states. Across all models, free lunch segregation and white-black segregation are both significant predictors of the ICCs; white-Hispanic segregation is not. The coefficient on free-lunch segregation is much larger than the coefficient on white-black segregation, particular in ELA (0.51 vs 0.13 in math; 0.61 vs 0.05 in ELA), indicating that between-district economic segregation may be a much more important driver of between-district test score variation than is racial segregation.

Figure 2 plots the Empirical Bayes ICC estimates against the state-average white-black and free lunch segregation to visually demonstrate their bivariate relationships. The correlations of white-black segregation with math and ELA ICCs are 0.77 and 0.69, respectively, while the correlations of free-lunch segregation and ICCS are even higher: 0.89 and 0.92, respectively. These strong correlations make clear that segregation is closely associated with the amount of between-district variation in the average student test scores among states.

[Figure 2]

Our last analysis investigates the association between segregation and the rate at which the ICC changes across grades. Table 3 shows that the rate of change of the ICC from grade 3 to 8 is positive on average, but varies significantly among states. We hypothesized that in states with more segregation the compounding effects of differential exposure to resources may be larger than in less segregated states because the contrast in resources among districts is likely starker. The regression estimates in Table 3 provide some support for this hypothesis. Although none of the three segregation measures is individually significant as a predictor of the growth of ICCs across grades, the three measures jointly predict ICCs ($p<.05$), and the signs of the coefficients on the segregation variables are generally positive, indicating that in states with higher levels of racial and socioeconomic segregation, ICCs grow slightly faster from grade 3 to 8, on average. Nonetheless, only a modest fraction (18-24%) of the variance in the grade slope is explained by the segregation and structural variables in Model 5, indicating that factors other than segregation play an important role in shaping changes over grades in between-district academic performance.

[Table 3]

Discussion

Our population-based analyses show that test scores vary substantially among states, ranging from an ICC near zero – no difference in average test scores among districts – to an ICC of 0.237 – average performance differs considerably among districts. It further confirms two patterns suggested by prior research: that between-district variation is larger in math than in ELA and, that between-district variance grows over grades. And, uncovers a new one, showing that between-district variation has grown over time. Specifically, we find that on average ICCs are 12% larger in math compared to ELA (although they are highly correlated), grow by up to 30% as students move from third through eighth grade, and have increased by more than 30% over the twelve cohorts in our sample.

Almost 90% of the test score variation among states can be accounted for by patterns of between-district white-black and economic segregation in addition to structural characteristics of school districts. In particular, states with high levels of white-black and economic segregation have, on average, more between-district variation. Further, this relationship is particularly strong for economic segregation. This is consistent with our hypothesis that segregation leads to large between-district differences in total family, neighborhood and school socioeconomic conditions and resources which, in turn, generates between-district test score variation.

Although our analyses show that the relationship between segregation and between-district variation is similar in math and ELA, the larger ICCs and higher growth rates in math suggest that mathematics test scores may be more sensitive than ELA scores to context. Evidence from prior research finds that educational interventions more often yield larger effects on test scores in mathematics than in ELA (e.g., Decker, Mayer, & Glaserman, 2004; Dobbie & Fryer, 2011; Jacob, 2005). Therefore, it may be that exposure to differential resources, particularly in the school context, may generate larger variability in mathematics test scores relative to ELA, and this difference may compound more over time.

About one fifth of the variance in the grade slopes among states is explained by segregation. ICCs grow slightly faster in more segregated states compared with others, suggesting that the effects of differences in exposure to resources may accumulate over time. However, most of the variation in the grade slopes among states is unexplained by segregation and structural factors.

ICCs have grown substantially—by 30%—over the 12 cohorts we examine here. Roughly speaking, that means the between-district standard deviation of test scores has grown by 15%, a sizeable change over a decade. Given the strong association we find between segregation and ICCs, the increase in ICCs may be driven by the trend of rising income segregation between school districts documented by Owens (2016). Owens, however, finds that between-district income segregation among families with children grew by about 10% from 2000 to 2010, substantially less than the 30% increase in ICCs we observe between the cohorts of children starting kindergarten from 2000 to 2011. This suggests that increasing income segregation alone may not explain the increasing ICCs. Future research should investigate the rapid growth in between-district variation in test scores.

Because we have no direct measures of school district quality or measures of the between-district test score variation at the start of formal schooling, our analyses here cannot distinguish the relative importance of family, neighborhood, and school factors in shaping patterns of between-district test score variation. Moreover, we believe that they are likely not fully separable in practice. If local socioeconomic conditions shape school quality—because affluent districts are able to marshal more economic, social, and political resources and to attract and retain more skilled teachers and staff—then a key channel through which local socioeconomic conditions shape educational outcomes is through their effects on school quality.

What is clear from this analysis, however, is the strong association between context—particularly socioeconomic context—and educational opportunity. We need to break this association in order to reduce educational inequality. Research should focus on disentangling the contribution of family background and school quality (to the extent possible) so that we can isolate what, if any, aspects of school quality drive between-district variation in academic success and how we can improve those in schools in lower-income, higher-minority communities.

References

Bayer, P., Ferreira, F., McMillan, R., Bajari, P., Berry, S., Black, S., … Staiger, D. (2004). Tiebout Sorting, Social Multipliers and the Demand for School Quality. *National Bureau of Economic Research Working Paper Series*, *10871*(203). https://doi.org/10.3386/w10871

Bayer, P., Ferreira, F., Mcmillan, R., Journal, S., August, N., & Bayer, P. (2007). A Unified Framework for Measuring Preferences for Schools and Neighborhoods. *Journal of Political Economy*, *115*(4), 588–638. Retrieved from http://www.jstor.org/stable/10.1086/522381

Bischoff, K. (2008). School District Fragmentation and Racial Residential Segregation: How Do Boundaries Matter? *Urban Affairs Review*, *44*(2), 182–217. https://doi.org/10.1177/1078087408320651

Bloom, H. S., Richburg-Hayes, L., & Black, A. R. (2007). Using Covariates to Improve Precision for Studies That Randomize Schools to Evaluate Educational Interventions. *Educational Evaluation and Policy Analysis*, *29*(1), 30–59. https://doi.org/10.3102/0162373707299550

Card, D., & Krueger, A. B. (1992). Does School Quality Matter ? Returns to Education and the Characteristics of Public Schools in the United States. *Journal of Political Economy*, *100*(1), 1–40. Retrieved from http://www.jstor.org/stable/2138804

Chetty, R., Friedman, J. N., Hilger, N., Saez, E., Schanzenbach, D. W., & Yagan, D. (2011). How does your kindergarten classroom affect your earnings? Evidence from project star. *Quarterly Journal of Economics*, *126*(4), 1593–1660. https://doi.org/10.1093/qje/qjr041

Decker, P., Mayer, D., & Glaserman, S. (2004). *The Effects of Teach for America on Students: Findings from a National Evaluation*.

Deming, D. J., Hastings, J. S., Kane, T. J., & Staiger, D. O. (2014). School Choice, School Quality, and Postsecondary Attainment. *American Economic Review*, *104*(3), 991–1013. https://doi.org/10.1257/aer.104.3.991

Dobbie, W., & Fryer, R. G. (2011). Are high-quality schools enough to increase achievement among the poor? Evidence from the Harlem Children’s Zone. *American Economic Journal: Applied Economics*, *3*(3), 158–187. https://doi.org/10.1257/app.3.3.158

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass Correlation Values for Planning Group-Randomized Trials in Education. *Educational Evaluation and Policy Analysis*, *29*(1), 60–87. https://doi.org/10.3102/0162373707299706

Hedges, L. V., & Hedberg, E. C. (2014). Intraclass Correlations and Covariate Outcome Correlations for Planning Two- and Three-Level Cluster-Randomized Experiments in Education. *Evaluation Review*, *37*(6), 445–489. https://doi.org/10.1177/0193841X14529126

Herfindahl, O. C. (1950). *Concentration in the U.S. Steel Industry.* Columbia University.

Hirschman, A. O. (1964). The Paternity of an Index. *American Economic Review*, *54*(5), 761.

Hirschman, A. O., & Albert O., H. (1945). *National Power and the Structure of Foreign Trade*. *National Power and the Structura of Foreign Trade*. Los Angeles: University of California Press.

Jacob, B. A. (2005). Accountability, incentives and behavior: The impact of high-stakes testing in the Chicago Public Schools. *Journal of Public Economics*, *89*(5–6), 761–796. https://doi.org/10.1016/j.jpubeco.2004.08.004

Jacob, R., Zhu, P., & Bloom, H. (2010). New Empirical Evidence for the Design of Group Randomized Trials in Education. *Journal of Research on Educational Effectiveness*, *3*(2), 157–198. https://doi.org/10.1080/19345741003592428

Konstantopoulos, S. (2009). Incorporating cost in power analysis for three-level cluster-randomized designs. *Evaluation Review*, *33*(4), 335–57. https://doi.org/10.1177/0193841X09337991

Konstantopoulos, S. (2011). A More Powerful Test in Three-Level Cluster Randomized Designs. *Journal of Research on Educational Effectiveness*, *4*(4), 354–369. https://doi.org/10.1080/19345747.2010.519824

Konstantopoulos, S. (2012). The Impact of Covariates on Statistical Power in Cluster Randomized Designs: Which Level Matters More? *Multivariate Behavioral Research*, *47*(3), 392–420. https://doi.org/10.1080/00273171.2012.673898

Lankford, H., Loeb, S., & Wykoff, J. (2002). Teacher Sorting and the Plight of Urban Schools: A Descriptive Analysis. *Educational Evaluation and Policy Analysis*, *24*(1), 37–62. https://doi.org/10.3102/01623737024001037

Magnuson, K. A., Meyers, M. K., Ruhm, C. J., & Waldfogel, J. (2004). Inequality in Preschool Education and School Readiness. *American Educational Research Journal Spring*, *41*(1), 115–157. https://doi.org/https://doi.org/10.3102/00028312041001115

Massey, D. S., & Denton, N. A. (1988). The Dimensions of Residential Segregation, *67*(2), 281–315. Retrieved from http://www.jstor.org/stable/2579183

Owens, A. (2016). Inequality in Children's Contexts: The Economic Segregation of Households With and Without Children. *American Sociological Review*, *81*(3), 1–26. https://doi.org/10.1177/0003122416642430

Raudenbush, S. W., Martinez, A., & Spybrook, J. (2007). Strategies for Improving Precision in Group-Randomized Experiments. *Educational Evaluation and Policy Analysis*, *29*(1), 5–29. https://doi.org/10.3102/0162373707299460

Reardon, S. F. (2011). The Widening Academic Achievement Gap Between the Rich and the Poor: New Evidence and Possible Explanations. In G. J. Duncan & R. J. Murnane (Eds.), *Whither Opportunity? Rising Inequality, Schools, and Children's Life Chances* (pp. 91–115). New York: Russell Sage Foundation.

Reardon, S. F., & Ho, A. D. (2015). *Practical Issues in Estimating Achievement Gaps From Coarsened Data*. *Journal of Educational and Behavioral Statistics* (Vol. 40). https://doi.org/10.3102/1076998615570944

Reardon, S. F., Shear, B. R., Castellano, K. E., & Ho, A. D. (2016). Using Heteroskedastic Ordered Probit Models to Recover Moments of Continuous Test Score Distributions

From Coarsened Data. *Journal of Educational and Behavioral Statistics*, *42*(16–2), 3–45. https://doi.org/10.3102/1076998616666279

Sampson, R. J., Sharkey, P., & Raudenbush, S. W. (2008). Durable effects of concentrated disadvantage on verbal ability among African-American children. *Proceedings of the National Academy of Sciences*, *105*(3), 845–852. https://doi.org/10.1073/pnas.0710189104

Schochet, P. Z. (2008). Statistical Power for Random Assignment Evaluations of Education Programs. *Journal of Educational and Behavioral Statistics*, *33*(1), 62–87. https://doi.org/10.3102/1076998607302714

Shear, B. R., & Reardon, S. F. (2016). HETOP: Stata module for estimating heteroskedastic ordered probit models with ordered frequency data. Retrieved from https://ideas.repec.org/c/boc/bocode/s458287.html

Sirin, S. R. (2005). Socioeconomic Status and Academic Achievement: A Meta-Analytic Review of Research. *Review of Educational Research*, *75*(3), 417–453. https://doi.org/10.3102/00346543075003417

Theil, H., & Finezza, A. J. (1971). A Note on the Measurement of Racial Integation of Schools by Means of Informational Concepts. *Journal of Mathematical Sociology*, (1), 187–94.

Tiebout, C. M. (1956). A Pure Theory of Local Expenditures. *Journal of Political Economy*, (64), 416–424.

Valentino, R. (2015). Will Public Pre-K Really Close Achievement Gaps? Gaps in prekindergarten quality between students and across states.

Westine, C. D., Spybrook, J., & Taylor, J. A. (2013). An empirical investigation of variance design parameters for planning cluster-randomized trials of science achievement. *Evaluation Review*, *37*(6), 490–519. https://doi.org/10.1177/0193841X14531584

Wodtke, G. T., Elwert, F., & Harding, D. J. (2016). Neighborhood Effect Heterogeneity by Family Income and Developmental Period. *American Journal of Sociology*, *121*(4), 1168–1222. https://doi.org/10.1086/684137

Zhu, P., Jacob, R., Bloom, H., & Xu, Z. (2012). Designing and Analyzing Studies That Randomize Schools to Estimate Intervention Effects on Student Academic Outcomes Without Classroom-Level Information. *Educational Evaluation and Policy Analysis*, *34*(1), 45–68. https://doi.org/10.3102/0162373711423786

Footnotes

[1] ICCs are a common measure of between-group variance. A substantial body of literature has investigated how to improve the estimation of between-*school* ICCs to inform experimental design using data from individual districts, state data, or national representative samples (e.g., Bloom, Richburg-Hayes, & Black, 2007; Hedges & Hedberg, 2007, 2014; Jacob et al., 2010; Konstantopoulos, 2011; 2012; Schochet, 2008; Westine, Spybrook, & Taylor, 2013; Zhu, Jacob, Bloom, & Xu, 2012). These studies generally focus on using covariates to explain variance, and thereby improve the minimal detectable effect size (MDES). We have not included a detailed review of this literature as it is beyond the scope of our paper. However, these studies underscore the relevance of using the ICC as a measure of between-district variation.

[2] Of course, not every public school student attends a school located in the geographic district in which he or she resides, but the overwhelming majority do.

[3] Note that if we had computed the between-school ICCs this would have been the case and the ICC estimates would be biased. Therefore, we do not report school-level ICCs.

[4] Specifically, we fit the homoskedastic model in 82 of the 3,795 state-grade-year-subject cases. These cases include Colorado (36 cases; all subjects and grades in the 2008-09, 2009-10, and 2010-11 school years), Florida (12 cases; all subjects and grades in the 2008-09 school year), New Mexico (12; all subjects and grades in the 2014-15 school year), South Carolina (12 cases; all subjects and grades in the 2010-11 school year), and

Texas (10 cases; all subjects, 3rd through 6th grades in the 2011-12 school year).

[5] The reliability data for each state's subject-grade-year tests were provided by Reardon and Ho (2015) and supplemented with additional publicly available information from state technical reports. For cases where no information was available, test reliabilities were imputed using data from other grades and years in the same state.

[6] The district fragmentation can be interpreted as the probability that two randomly chosen students in a state are enrolled in different school districts. For state s, grade g, and year y, it is defined as: $HHI_{sgy} = \sum_{d \in s} \left(\frac{T_{dgy}}{T_{sgy}}\right)\left(1 - \frac{T_{dgy}}{T_{sgy}}\right)$, where T_{dgy} and T_{sgy} are number of students in a given grade (g) and year (y) enrolled in district d or state s, respectively.

[7] Data files can be found on the CCD data page of the CCD website: http://nces.ed.gov/ccd/ccddata.asp.

[8] We considered other measures including the standard deviation of district enrollment and the coefficient of variation of district enrollment, but excluded these from our final models for parsimony, as they added no additional explanatory power.

[9] Because there is missing free lunch data in the CCD, we use multiple imputation at the school level. We impute missing data using free lunch and racial composition data from other grades and years in the same school. We then collapse the imputed school-level data to the district-level and estimate the between-district poor-non poor segregation.

[10] Note that the larger math ICCs are not a function of greater reliability of math

tests, since the reliability of state math tests and ELA tests do not differ appreciably and we adjust for reliability.

Tables and Figures

Table I: State Average Structural and Segregation Covariates

Variable	Mean	Standard Deviation	Min	Max
Number of Districts	265.21	(238.58)	15.64	1026.90
Natural Log of Number of Districts	5.17	(0.99)	2.75	6.93
Mean District Grade-Level Enrollment	435.07	(573.38)	38.22	2823.49
Natural Log of Mean District Grade-Level Enrollment	5.57	(0.97)	3.64	7.95
District Fragmentation (HHI)	0.95	(0.08)	0.43	0.99
Transformed District Fragmentation (ln(1/(1-HHI)))	3.55	(0.92)	0.57	5.13
Between-District, Within State White/Black Segregation (H)	0.26	(0.14)	0.06	0.52
Between-District, Within State White/Hispanic Segregation (H)	0.19	(0.11)	0.03	0.45
Between-District, Within State Free Lunch Segregation (H)	0.12	(0.07)	0.01	0.31

Note: Summary statistics include one observation for each of the 49 states included in the subsequent analyses, which is the average of the variable across grades (3-8) & years (2009-2015). All log transformations are natural logs.

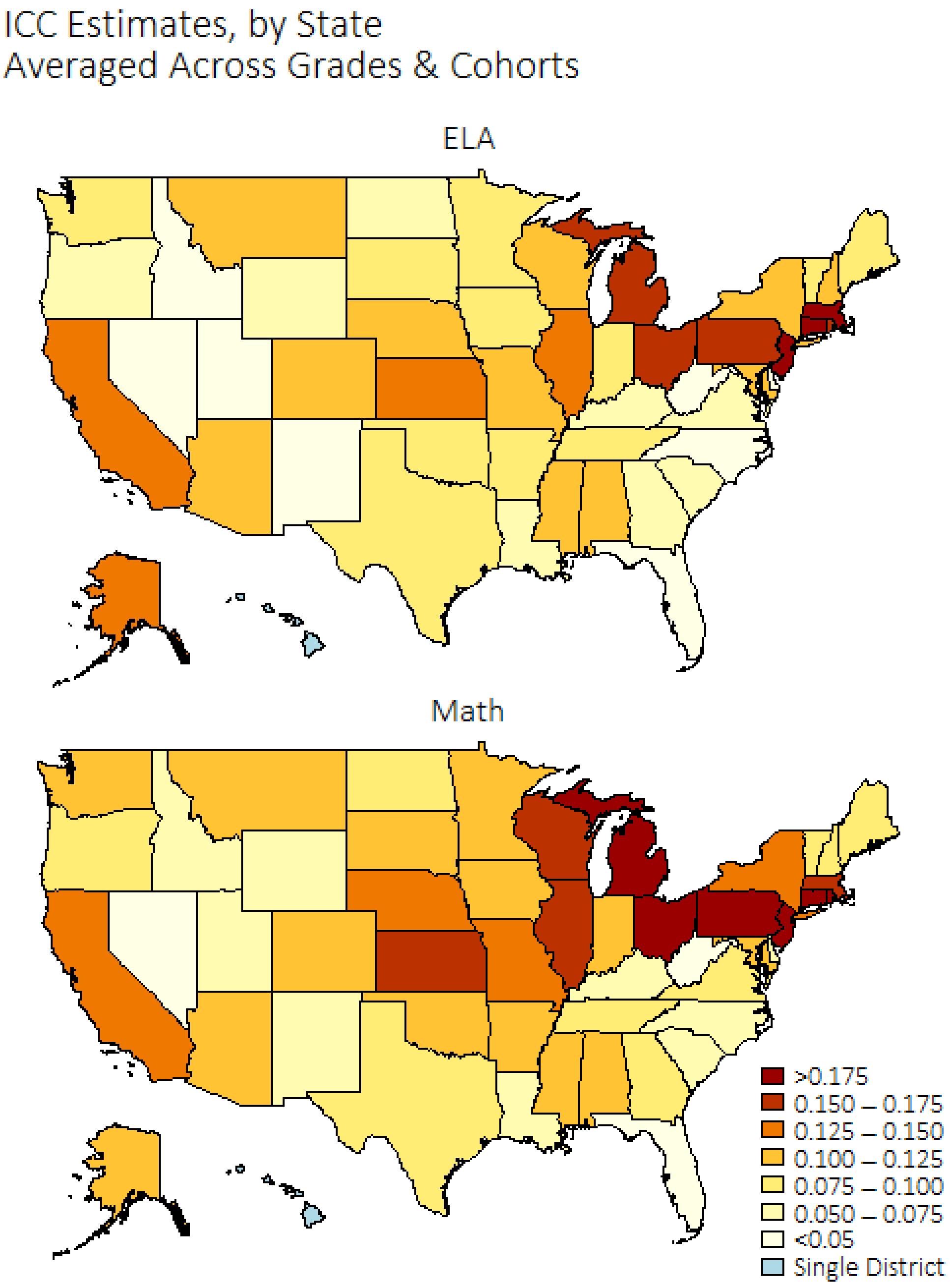

Figure I: Maps of ICC Estimates by State and Subject Averaged Across Grades & Years

Table II: Multivariate Relationships Among State Intraclass Correlations and Measures of Between-District Segregation

	Math				ELA			
	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
Intercept	0.0994 ***	0.0994 ***	0.0993 ***	0.0992 ***	0.0882 ***	0.0882 ***	0.0881 ***	0.0882 ***
	(0.0072)	(0.0157)	(0.0036)	(0.0034)	(0.0066)	(0.0201)	(0.0031)	(0.0033)
2011 Cohort	0.0298 ***	0.0298 ***	0.03 ***	0.0300 ***	0.0309 ***	0.0310 ***	0.031 ***	0.0310 ***
	(0.0060)	(0.0065)	(0.0062)	(0.0062)	(0.0055)	(0.0085)	(0.0061)	(0.0054)
Grade	0.0063 ***	0.0063 ***	0.0063 ***	0.0063 ***	0.0038 ***	0.0038 **	0.0038 ***	0.0038 ***
	(0.0008)	(0.0010)	(0.0008)	(0.0008)	(0.0006)	(0.0012)	(0.0007)	(0.0006)
Ln Number of Districts		0.0124		-0.0097 *		0.0145		-0.0100 *
		(0.0169)		(0.0039)		(0.0173)		(0.0051)
Ln Mean Enrollment		-0.0116		-0.0117 *		0.0002		-0.0051
		(0.0335)		(0.0049)		(0.0481)		(0.0054)
Transformed Herfindahl Index		0.0112		0.0051		0.0108		0.0065
		(0.0118)		(0.0058)		(0.0265)		(0.0090)
White-Black Segregation			0.1206 ***	0.1301 **			0.0368	0.0457
			(0.0347)	(0.0422)			(0.0241)	(0.0438)
White-Hispanic Segregation			-0.126 *	-0.0605			-0.0668	-0.0170
			(0.0583)	(0.0876)			(0.0427)	(0.0873)
Free Lunch Segregation			0.6032 ***	0.5054 ***			0.6535 ***	0.6085 ***
			(0.0732)	(0.1186)			(0.0725)	(0.1132)
Within-State SD	0.0120	0.0120	0.0120	0.0120	0.0102	0.0102	0.0102	0.0102
Between-State Intercept SD	0.0482	0.0386	0.0191	0.0166	0.0471	0.0402	0.0177	0.0167
Between-State Grade SD	0.0050	0.0050	0.0051	0.0050	0.0043	0.0043	0.0043	0.0043
Between-State Cohort SD	0.0034	0.0034	0.0034	0.0033	0.0030	0.0030	0.0030	0.0030
Reliability - Intercept	0.998	0.997	0.987	0.983	0.999	0.998	0.990	0.988
Reliability - Grade	0.897	0.898	0.898	0.898	0.904	0.904	0.904	0.904
Reliability - Cohort	0.902	0.903	0.901	0.901	0.911	0.911	0.911	0.911
P-value from the joint hypothesis test that all the structural covariates are jointly equal to zero		0.16		0.03		0.47		0.28
P-value from the joint hypothesis test that all the segregation measures are jointly equal to zero			0.00	0.00			0.00	0.00
R^2 (Relative to Model (1))		0.36	0.84	0.88		0.27	0.86	0.87

+ p<0.10 * p<0.05 ** p<0.01 *** p<0.001; Standard errors in parentheses.

Notes: The ELA models have 1923 state-grade-year observations clustered in 49 states and the math models have 1866 state-grade-year observations clustered in 49 states. Model (1) is the baseline random coefficient model. It includes an intercept (corresponding to the average ICC for the 2000 cohort in grade 5.5), cohort dummies for 2001 through 2011 (the coefficients on the dummies for 2001 to 2010 have been omitted from this table for parsimony), a linear grade term centered at 5.5, a random coefficient on the centered grade term, and a random coefficient on a linear cohort term centered at 2005.5. Model (2) adds structural controls to Model (1). Model (3) adds segregation measures to Model (1).Model (4) adds both stuctural controls and segregation measures to Model (1). Structural controls include: the natural log number of districts, the natural log mean enrollment, and the transformed Herfindahl index (ln(1/(1-HHI). Segregation measures include: white-black segregation, white-Hispanic segregation, and free lunch segregation. All covariates are grand mean centered.

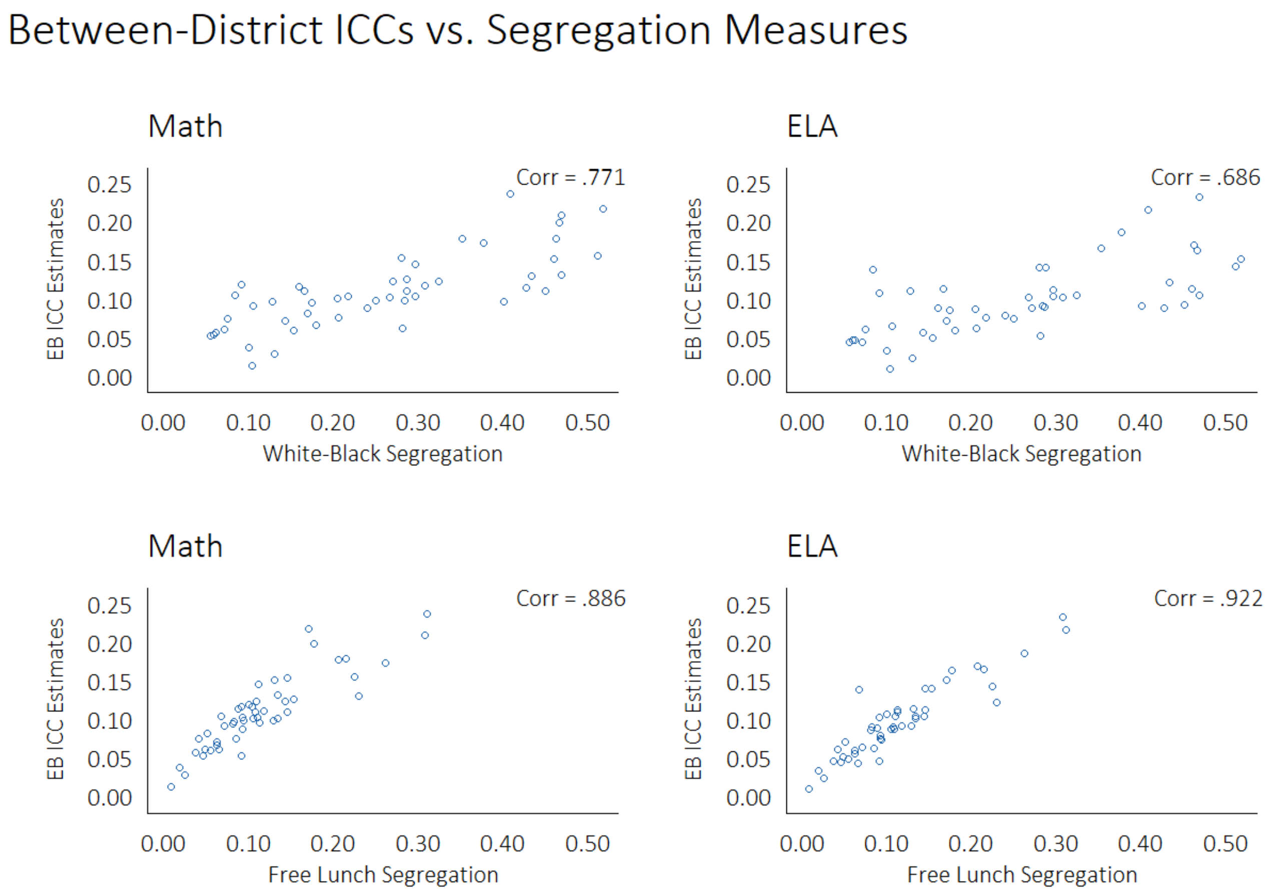

Figure II: Between District ICCs vs. Segregation

Table III: Explaining Variation in Grade Trends using District Structure and Between-District Segregation

	Math		ELA	
	(4)	(5)	(4)	(5)
Intercept	0.0992 ***	0.0992 ***	0.0882 ***	0.0882 ***
	(0.0034)	(0.0035)	(0.0033)	(0.0030)
2011 Cohort	0.03 ***	0.0299 ***	0.031 ***	0.031 ***
	(0.0062)	(0.0060)	(0.0054)	(0.0054)
Grade	0.0063 ***	0.0063 ***	0.0038 ***	0.0038 ***
	(0.0008)	(0.0007)	(0.0006)	(0.0006)
Log Number of Districts - X - Grade		-0.0009		-0.0017 *
		(0.0009)		(0.0008)
Log Mean Enrollment - X - Grade		0.0013 +		-0.0004
		(0.0007)		(0.0006)
Herfindahl Index - X - Grade		0.001		0.0013 +
		(0.0009)		(0.0007)
White-Black Segregation - X - Grade		0.0092		-0.0024
		(0.0067)		(0.0057)
White-Hispanic Segregation - X - Grade		0.0099		0.0174 +
		(0.0107)		(0.0091)
Free Lunch Segregation - X - Grade		-0.0006		0.0013
		(0.0165)		(0.0141)
Within-State SD	0.0120	0.0120	0.0102	0.0102
Between-State Intercept SD	0.0166	0.0166	0.0167	0.0167
Between-State Grade SD	0.0050	0.0044	0.0043	0.0039
Between-State Cohort SD	0.0033	0.0033	0.0030	0.0030
Reliability - Intercept	0.983	0.983	0.988	0.988
Reliability - Grade	0.898	0.869	0.904	0.886
Reliability - Cohort	0.901	0.901	0.911	0.911
P-value from the joint hypothesis test that all the structural covariate interactions with grade are jointly equal to zero		0.05		0.15
P-value from the joint hypothesis test that all the segregation measure interactions with grade are jointly equal to zero		0.03		0.02
R^2 (Relative to Model (3))		0.24		0.18

+ p<0.10 * p<0.05 ** p<0.01 *** p<0.001; Standard errors in parentheses.

Notes: The ELA models have 1923 state-grade-year observations clustered in 49 states and the math models have 1866 state-grade-year observations clustered in 49 states. Model (4) is the same as shown in Table II. It is the random coefficient model including both structural controls and segregation measures. Model (5) adds interaction terms between grade and the structural controls and segregation measures. Structural controls include: the natural log number of districts, the natural log mean enrollment, and the transformed Herfindahl index (ln(1/(1-HHI). Segregation measures include: white-black segregation, white-Hispanic segregation, and free lunch segregation. All covariates are grand mean centered.

Appendix Tables

Appendix Table I: Summary of State ICC Estimates, Structural Covariates and Segregation Measures

State	Number of Grade-Year-Subject Observations	ELA ICC Estimate	Math ICC Estimate	Number of Districts	Mean Enrollment	Herfindahl Index	White-Black Segregation	White-Hispanic Segregation	Free Lunch Segregation
Alabama	80	0.102	0.117	133.2	429.3	0.978	0.126	0.309	0.092
Alaska	84	0.138	0.104	53.4	179.8	0.806	0.072	0.085	0.068
Arizona	84	0.113	0.110	196.3	423.1	0.976	0.240	0.167	0.147
Arkansas	84	0.088	0.114	243.1	149.9	0.985	0.218	0.428	0.089
California	60	0.140	0.126	913.5	511.0	0.985	0.262	0.288	0.154
Colorado	74	0.102	0.102	179.8	350.3	0.957	0.169	0.268	0.135
Connecticut	72	0.216	0.237	161.1	256.3	0.986	0.340	0.409	0.312
Delaware	84	0.046	0.057	15.6	626.7	0.908	0.072	0.063	0.038
Florida	78	0.024	0.028	72.6	2823.5	0.947	0.213	0.131	0.026
Georgia	84	0.074	0.099	183.2	703.3	0.971	0.165	0.251	0.094
Idaho	72	0.045	0.053	114.8	187.1	0.955	0.132	0.056	0.047
Illinois	82	0.143	0.156	809.8	189.3	0.960	0.409	0.513	0.226
Indiana	84	0.092	0.110	296.8	268.0	0.991	0.237	0.452	0.119
Iowa	84	0.075	0.103	338.1	104.8	0.987	0.201	0.218	0.094
Kansas	72	0.140	0.153	290.4	121.8	0.973	0.251	0.281	0.146
Kentucky	84	0.052	0.062	176.3	287.0	0.968	0.116	0.282	0.050
Louisiana	84	0.060	0.067	78.4	681.5	0.967	0.134	0.181	0.063
Maine	79	0.079	0.088	173.7	79.6	0.987	0.074	0.241	0.093
Maryland	76	0.103	0.103	24.4	2556.4	0.905	0.236	0.298	0.111
Massachusetts	84	0.186	0.173	264.7	271.4	0.989	0.389	0.378	0.263
Michigan	84	0.152	0.217	581.2	200.2	0.991	0.244	0.519	0.172
Minnesota	84	0.090	0.110	354.5	174.5	0.985	0.176	0.287	0.109
Mississippi	84	0.105	0.124	150.5	249.1	0.983	0.129	0.325	0.145
Missouri	84	0.105	0.132	523.2	129.9	0.990	0.199	0.470	0.135
Montana	60	0.107	0.119	267.4	40.5	0.971	0.067	0.092	0.101
Nebraska	66	0.112	0.145	250.5	87.3	0.944	0.229	0.298	0.113
Nevada	60	0.009	0.013	18.7	1905.6	0.433	0.033	0.105	0.009
New Hampshire	82	0.110	0.096	141.6	103.5	0.980	0.173	0.128	0.114
New Jersey	72	0.232	0.209	497.5	199.5	0.994	0.419	0.470	0.309
New Mexico	84	0.046	0.053	92.1	273.6	0.891	0.090	0.060	0.093
New York	66	0.122	0.130	682.2	290.2	0.866	0.414	0.434	0.230
North Carolina	84	0.049	0.059	116.0	995.2	0.966	0.067	0.155	0.056
North Dakota	73	0.064	0.091	168.1	43.5	0.953	0.156	0.107	0.072
Ohio	84	0.164	0.199	615.7	212.7	0.994	0.209	0.468	0.178
Oklahoma	83	0.088	0.123	523.0	91.9	0.983	0.207	0.272	0.109
Oregon	72	0.071	0.082	191.3	222.3	0.972	0.106	0.171	0.052
Pennsylvania	84	0.170	0.178	500.6	260.8	0.985	0.384	0.463	0.207
Rhode Island	77	0.165	0.178	36.0	295.0	0.940	0.453	0.354	0.216
South Carolina	84	0.056	0.071	87.1	632.5	0.968	0.060	0.144	0.064
South Dakota	72	0.088	0.117	153.6	61.6	0.950	0.090	0.161	0.106
Tennessee	84	0.091	0.097	136.9	543.8	0.966	0.167	0.402	0.084
Texas	76	0.092	0.098	1026.9	361.4	0.991	0.298	0.285	0.130
Utah	84	0.044	0.061	43.0	1085.3	0.923	0.105	0.072	0.066
Vermont	72	0.086	0.094	174.3	38.2	0.987	0.079	0.175	0.082
Virginia	70	0.062	0.076	133.6	702.9	0.959	0.143	0.207	0.086
Washington	60	0.087	0.101	288.8	271.1	0.986	0.198	0.206	0.106
West Virginia	84	0.033	0.037	56.9	361.0	0.965	0.115	0.101	0.020
Wisconsin	84	0.114	0.151	415.6	146.6	0.984	0.227	0.461	0.132
Wyoming	57	0.061	0.075	49.5	138.6	0.935	0.066	0.075	0.043
Average	***77.4***	***0.098***	***0.110***	***265.2***	***435.1***	***0.951***	***0.191***	***0.260***	***0.116***

Notes: The average Empirical Bayes ICC estimate is estimated from Model (1). It is the average across all cohorts (2000-2011) in the average grade (grade = 5.5). The covariate values are averages across the grades and years in our sample.

Appendix Table II: Multivariate Relationships Among State Intraclass Correlations and Measures of Between-District Segregation - All Coefficients

	Math				ELA			
	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
Intercept	0.0994 ***	0.0994 ***	0.0993 ***	0.0992 ***	0.0882 ***	0.0882 ***	0.0881 ***	0.0882 ***
	(0.0072)	(0.0157)	(0.0036)	(0.0034)	(0.0066)	(0.0201)	(0.0031)	(0.0033)
2001 Cohort	0.0015	0.0015	0.0015	0.0015	0.0012	0.0012	0.0012	0.0012
	(0.0018)	(0.0018)	(0.0018)	(0.0018)	(0.0013)	(0.0014)	(0.0013)	(0.0013)
2002 Cohort	0.0036	0.0035	0.0036	0.0036	0.0037 **	0.0037 *	0.0037 *	0.0037 **
	(0.0023)	(0.0023)	(0.0023)	(0.0023)	(0.0014)	(0.0018)	(0.0015)	(0.0014)
2003 Cohort	0.0045 +	0.0045 +	0.0045 +	0.0045 +	0.004 *	0.004	0.004 *	0.004 *
	(0.0025)	(0.0025)	(0.0025)	(0.0025)	(0.0018)	(0.0025)	(0.0020)	(0.0018)
2004 Cohort	0.0046 +	0.0046 +	0.0047 +	0.0047 +	0.0042 *	0.0042	0.0042 *	0.0042 *
	(0.0026)	(0.0026)	(0.0026)	(0.0026)	(0.0019)	(0.0029)	(0.0021)	(0.0020)
2005 Cohort	0.0074 *	0.0074 *	0.0075 *	0.0076 *	0.0061 **	0.0061 +	0.0062 *	0.0062 **
	(0.0029)	(0.0030)	(0.0030)	(0.0030)	(0.0021)	(0.0035)	(0.0024)	(0.0021)
2006 Cohort	0.0088 **	0.0088 **	0.0089 **	0.0089 **	0.0066 *	0.0066	0.0067 *	0.0067 *
	(0.0032)	(0.0034)	(0.0033)	(0.0034)	(0.0026)	(0.0043)	(0.0030)	(0.0027)
2007 Cohort	0.0106 **	0.0106 **	0.0108 **	0.0108 **	0.0095 ***	0.0095 *	0.0096 **	0.0096 **
	(0.0035)	(0.0037)	(0.0036)	(0.0037)	(0.0028)	(0.0048)	(0.0033)	(0.0030)
2008 Cohort	0.0149 **	0.0149 **	0.015 **	0.0150 **	0.0122 ***	0.0122 *	0.0123 **	0.0123 ***
	(0.0047)	(0.0048)	(0.0048)	(0.0049)	(0.0035)	(0.0057)	(0.0041)	(0.0036)
2009 Cohort	0.0190 ***	0.0189 ***	0.0191 ***	0.0191 ***	0.0151 ***	0.0151 *	0.0152 ***	0.0152 ***
	(0.0049)	(0.0051)	(0.0050)	(0.0051)	(0.0038)	(0.0063)	(0.0044)	(0.0039)
2010 Cohort	0.0252 ***	0.0252 ***	0.0254 ***	0.0254 ***	0.0218 ***	0.0218 **	0.0219 ***	0.0219 ***
	(0.0052)	(0.0056)	(0.0054)	(0.0055)	(0.0046)	(0.0074)	(0.0052)	(0.0046)
2011 Cohort	0.0298 ***	0.0298 ***	0.03 ***	0.0300 ***	0.0309 ***	0.0310 ***	0.031 ***	0.0310 ***
	(0.0060)	(0.0065)	(0.0062)	(0.0062)	(0.0055)	(0.0085)	(0.0061)	(0.0054)
Grade	0.0063 ***	0.0063 ***	0.0063 ***	0.0063 ***	0.0038 ***	0.0038 **	0.0038 ***	0.0038 ***
	(0.0008)	(0.0010)	(0.0008)	(0.0008)	(0.0006)	(0.0012)	(0.0007)	(0.0006)
Ln Number of Districts		0.0124		-0.0097 *		0.0145		-0.0100 +
		(0.0169)		(0.0039)		(0.0173)		(0.0051)
Ln Mean Enrollment		-0.0116		-0.0117 *		0.0002		-0.0051
		(0.0335)		(0.0049)		(0.0481)		(0.0054)
Transformed Herfindahl Index		0.0112		0.0051		0.0108		0.0065
		(0.0118)		(0.0058)		(0.0265)		(0.0090)
White-Black Segregation			0.1206 ***	0.1301 **			0.0368	0.0457
			(0.0347)	(0.0422)			(0.0241)	(0.0438)
White-Hispanic Segregation			-0.126 *	-0.0605			-0.0668	-0.0170
			(0.0583)	(0.0876)			(0.0427)	(0.0873)
Free Lunch Segregation			0.6032 ***	0.5054 ***			0.6535 ***	0.6085 ***
			(0.0732)	(0.1186)			(0.0725)	(0.1132)
Within-State SD	0.0120	0.0120	0.0120	0.0120	0.0102	0.0102	0.0102	0.0102
Between-State Intercept SD	0.0482	0.0386	0.0191	0.0166	0.0471	0.0402	0.0177	0.0167
Between-State Grade SD	0.0050	0.0050	0.0051	0.0050	0.0043	0.0043	0.0043	0.0043
Between-State Cohort SD	0.0034	0.0034	0.0034	0.0033	0.0030	0.0030	0.0030	0.0030
Reliability - Intercept	0.998	0.997	0.987	0.983	0.999	0.998	0.990	0.988
Reliability - Grade	0.897	0.898	0.898	0.898	0.904	0.904	0.904	0.904
Reliability - Cohort	0.902	0.903	0.901	0.901	0.911	0.911	0.911	0.911
P-value from the joint hypothesis test that all the structural covariates are jointly equal to zero		0.16		0.03		0.47		0.28
P-value from the joint hypothesis test that all the segregation measures are jointly equal to zero			0.00	0.00			0.00	0.00
R^2 (Relative to Model (1))		0.36	0.84	0.88		0.27	0.86	0.87

+ p<0.10 * p<0.05 ** p<0.01 *** p<0.001; Standard errors in parentheses.

Notes: The ELA models have 1923 state-grade-year observations clustered in 49 states and the math models have 1866 state-grade-year observations clustered in 49 states. Model (1) is the baseline random coefficient model. It includes an intercept (corresponding to the average ICC for the 2000 cohort in grade 5.5), cohort dummies for 2001 through 2011, a linear grade term centered at 5.5, a random coefficient on the centered grade term, and a random coefficient on a linear cohort term centered at 2005.5. Model (2) adds structural controls to Model (1). Model (3) adds segregation measures to Model (1).Model (4) adds both stuctural controls and segregation measures to Model (1). Structural controls include: the natural log number of districts, the natural log mean enrollment, and the transformed Herfindahl index (ln(1/(1-HHI). Segregation measures include: white-black segregation, white-Hispanic segregation, and free lunch segregation. All covariates are grand mean centered.