Determinants of Between-School Variation in Student Achievement: Results from US Population Data

Benjamin R. Shear[1], Joseph Taylor[2], and Erin M. Fahle[3]

[1] University of Colorado Boulder

[2] University of Colorado Colorado Springs

[3] Stanford University

Author Note

Benjamin R. Shear https://orcid.org/0000-0002-9236-2927

Erin M. Fahle https://orcid.org/0000-0003-2404-6399

The authors acknowledge helpful comments from sean reardon on an earlier draft of the article. Correspondence concerning this article should be addressed to Benjamin R. Shear, School of Education, University of Colorado Boulder, 249 UCB, Boulder, CO 80309.

Email: benjamin.shear@colorado.edu

This manuscript is the version accepted for publication in *Journal of Research on Educational Effectiveness*. Final published version available at:

Shear, B. R., Taylor, J. A., & Fahle, E. M. (2026). Determinants of between-school variation in student achievement: Results from US population data. *Journal of Research on Educational Effectiveness*, 1–25. https://doi.org/10.1080/19345747.2025.2598314

Abstract

This paper uses school-level aggregate achievement data to estimate the intraclass correlation coefficient (ICC) – proportion of variation in test scores that is between schools within states – and how this is related to school characteristics. This study contributes to the literature by providing an up-to-date, comprehensive set of between-school ICC estimates based on population-level data and identifying relevant school context characteristics that can explain differences in ICC estimates. We estimate 6,021 state-grade-year-subject ICCs for grades 3-8 mathematics and reading/language arts (RLA) using state assessment data from 2009-2019 across all 50 states and the District of Columbia. The average math ICC is 0.194 and the average RLA ICC is 0.168. Average ICC estimates vary widely across states. The majority of between-state differences in ICC estimates (70% in math and 78% in RLA) can be explained by a small number of variables representing school structures and the level of between-school racial and economic segregation. We also find average ICC estimates increased over time, with larger increases in elementary grades and most of the increases occurring in 2015 or later.

Keywords: intraclass correlation; between-school variation; achievement; segregation; hierarchical linear modeling

Accurate, current estimates of variation in achievement levels among schools are critical for education research. The between-school intraclass correlation coefficient (ICC) of standardized test scores quantifies the proportion of test score variance due to differences in average test scores among schools relative to the total test score variance. The ICC provides essential information for planning cluster randomized trials (CRT) as sampling error and statistical power depend directly on the ICC of the relevant outcome variable (Hedges & Hedberg, 2007; Schochet, 2008). Knowing the ICC among a sample of schools recruited for a CRT from multiple states or districts is important for anticipating the required sample sizes or minimum detectable effect size of the CRT but may not convey substantively useful information. However, when the between-school ICC is estimated for a policy-relevant population of schools, the ICC itself is of substantive interest. For example, the ICC among all schools in a state conveys information about the distribution of educational opportunities across the state. Between-school ICCs are driven in part by differences in educational opportunities provided by schools and thus larger ICCs suggest greater inequality of educational opportunities provided to students attending different schools in the state. Comparing changes in between-school ICCs within a state across grades and years or across states provides a better understanding of educational opportunities.

To date, most studies have focused on generating ICC estimates for the purpose of study design (e.g., Hedges & Hedberg, 2007, 2014; Westine et al., 2013). Many of the ICC estimates in these papers are based on older data or are specific to a small number of states and years, and there have been few opportunities to validate the ICCs produced using both sample and population data. Moreover, only a small body of work has sought to quantify, across all states nationally, the relationship between school-level ICCs and school characteristics to inform our understanding of factors that may lead to greater test score variation between schools. A recent analysis of between-district ICCs using data from 49 states from 2009-2015 documented systematic associations between the magnitude of between-district ICCs and the context of schooling including the level of between-district racial and economic segregation (Fahle & Reardon, 2018). The lack of analogous evidence about between-school ICCs motivates the need for similar analyses of between-school ICC estimates from recent data. These analyses would ensure study designers have accurate information as they plan their work and provide scholars with a foundation to explore inequities among schools.

Our work extends prior literature by providing within-state between-school ICC estimates based on a recent, comprehensive national dataset of grade 3-8 math and reading/language arts (RLA) test scores. By producing between-school ICC estimates using a consistent methodology and recent state assessment data across all 50 states plus DC, the results here provide up-to-date estimates of between-school ICCs and allow us to investigate the extent to which between-school ICCs are correlated with the context of schooling across states. The scope of our analyses also allow for comparison to prior empirical ICC estimates to understand how different data (e.g., population vs. sample) and estimation methods affect ICC estimates.

Background

Many prior studies have reported between-school ICCs of math and RLA test scores across different grades, levels, and contexts across the United States to inform study planning (Bloom et al., 2007; Brandon et al., 2013; Hedberg & Hedges, 2014; Hedges & Hedberg, 2007, 2014; Jacob et al., 2010; Schochet, 2008; Shen et al., 2023; Westine et al., 2013; Zhu et al., 2012).[1,2] In one of the most widely cited studies reporting between-school ICC estimates, Hedges and Hedberg (2007) report between-school ICC estimates for math and reading tests in grades K-12 based on national probability samples. Their study provides a useful comparison set of estimates, although many of the datasets were from studies conducted in the 1980's and 1990's and may no longer represent the US population. Hedges and Hedberg (2007) report that in grades 3-8 the unconditional between-school ICC estimates ranged from 0.185 to 0.264 in math (M=0.222) and 0.174 to 0.271 in RLA (M=0.234) and were smaller in later grades compared to earlier ones in both subjects.

Two studies provide within-state between-school ICC estimates for state assessments based on population data (Hedges & Hedberg, 2014; Westine et al., 2013). Hedges and Hedberg (2014) estimate between-school math and RLA ICCs using population data from eight states in 2009-10 (AR, AZ, CO, KS, KY, MA, NC, and WI) and one state each in 2006-07 (FL), 2012-13 (LA), and 2011-12 (WV). The estimated ICCs for grades 3-8 varied significantly across states and were higher than the national estimates reported by Hedges and Hedberg (2007) in some states and lower in others. Average ICCs were higher in math, ranging from 0.037 to 0.398 (M=0.188), than in RLA, ranging from 0.035 to 0.348 (M=0.168). ICCs were on average larger in later grades, but the trends varied across states, with ICCs increasing across grades in some states and declining across grades in others. Westine et al. estimate between-school ICCs in Texas for grades 5, 8, 10, and 11 math, RLA, and science achievement tests administered from 2007-2011. In math the between-school ICC estimates were 0.168 and 0.163 in grades 5 and 8, respectively, while in RLA they were 0.156 and 0.099.

[1] Most prior studies report results for math and reading test scores; we use the more general term RLA to include reading tests and tests of similar constructs.

[2] More recently studies have presented estimates between-school ICCs of non-test score outcomes (Dong et al., 2016; Gray et al., 2016; Juras, 2016; Shen et al., 2023). These are beyond the scope of the current study.

The primary purpose of the aforementioned studies has been to provide ICCs for study design planning. While these studies document substantial variability in between-school ICCs across states included in the analyses, little is known about the sources of these differences, how they compare in other states, or how they vary over time. Three studies provide some context to anticipate this variability.

First, Fahle and Reardon (2018) provide a comprehensive analysis of within-state between-district ICCs, using population-level data for state assessments in grades 3-8 from 2009-2015 for 49 states (excluding HI, which comprises a single school district). While Fahle and Reardon focus on between-district ICCs, their findings are relevant because between-school ICCs include between-district variance. There was considerable variability of between-district ICCs during this period, with values ranging from near 0 (0.009 in RLA and 0.013 in math) to 0.232 in RLA and 0.237 in math. The between-district ICCs were larger on average in math than in RLA, increased on average from grades 3 to 8, and increased across cohorts from the 2000 cohort (students entering kindergarten in fall 2000) to the 2011 cohort. When adjusting for district structures (the number of districts, average district enrollment, and district fragmentation), racial and economic segregation explained the vast majority of between-state variance in ICC estimates. Segregation measures explained 84% and 86% of the between-state variation in average ICCs in math and RLA, respectively, while including district structure variables increased variance explained to 88% and 87%, respectively. These results were consistent with their hypothesis that racial and economic segregation leads to differences in educational opportunities reflected in larger between-district ICCs, and that these differences grow across grades. Based on this evidence, we would predict that states with more highly segregated school districts will have higher between-school ICCs. However, between-school

ICCs are also sensitive to within-district variation and thus associations with contextual factors may differ meaningfully.

Hedberg and Hedges (2014) analyzed the within-district between-school ICCs from the 11 states that were included in their prior analyses of within-state ICCs described above (Hedges & Hedberg, 2014). Hedberg and Hedges report that within-district between-school ICCs were higher in larger school districts (districts with more schools). In these analyses “large districts” were those with about 10 schools or more (depending upon grade level), and thus we do not necessarily expect the same associations to hold for within-state ICCs, where the distribution of schools within states differs substantially from within districts. That is, because most states have hundreds or even thousands of schools rather than 10, it is not clear whether the same association between the number of schools and ICCs will hold. Hedberg and Hedges also found that, conditional on district size, the within-district ICCs varied across grades, urbanicity, and free or reduced-price lunch eligibility rates, although the patterns were inconsistent. These results suggest the context of schooling within districts is related to ICCs, but do not provide clear predictions about how the statewide context of schooling influences between-school ICCs.

One recent study reported preliminary estimates of between-school within-state ICC estimates based on samples of schools participating in the 2022 National Assessment of Educational Progress (NAEP; Dogan et al., 2024). Between-school ICCs among grade 4 and 8 math and reading scores varied considerably across states and tended to be higher in math than in reading and in grade 4 than in grade 8. These between-school ICC estimates across states were weakly to moderately correlated with average state NAEP scores and school enrollments and were strongly positively correlated with measures of White-minority racial segregation and poverty segregation, each measured by the two-group normalized exposure index. The NAEP results further support the theme of inconsistent patterns in predicting ICCs from student or context characteristics, although more evidence is needed given they are based on a single year of data and a sample of schools.

Consistent with Fahle and Reardon (2018), we conceptualize average school test scores as representing students' overall educational opportunities provided by both in and out of school experiences. Based on this and prior work summarized above, we hypothesize three primary factors that could affect the magnitude of between-school ICCs across states, grades, and years: differences in in-school learning opportunities, differences in out-of-school learning opportunities, and differences in school structures. First, if some schools provide more learning opportunities, for example by implementing more effective curricula or hiring more effective teachers, then we would expect achievement to be higher in some schools leading to higher ICCs. Prior research on school and teacher value-added suggests that although these measures are imperfect, there are meaningful differences across schools and teachers in their effect on student test scores (e.g., Angrist et al., 2017; Chetty et al., 2014; Koedel et al., 2015).

Second, selection may affect between-school ICCs because families can often choose which school to enroll their children in either through their choice of neighborhood or via open enrollment policies, magnet schools, and other school choice policies. Higher ICCs may result from this selection if children with differential access to resources and educational opportunities outside of school are enrolling in different schools. This source of differences in ICCs is not directly caused by differences in school-related factors, although school-related factors may contribute to family choices. We expect selection effects to be stronger in places with smaller, more densely concentrated schools that provide greater opportunity for families to engage in school choice (e.g., more urban areas) and that higher levels of school racial and economic segregation reflect stronger selection effects, based on the well-documented association between student test scores and student family socioeconomic status and racial identity (e.g., Reardon, 2011; Reardon et al., 2015).

Third, the structure of schools, including the total number of schools and their average size, may affect ICCs. In the U.S., for example, elementary schools tend to be smaller with neighborhood-based enrollments, while middle and high schools enroll students from larger geographic areas. All else equal, we might expect places with fewer, larger schools enrolling students from larger geographic areas to enroll more heterogeneous populations, resulting in lower between-school ICCs. Differences in ICCs based on school structures are consistent with prior studies reviewed above, although these studies do not provide enough information to make precise predictions about how school structures statewide will impact between-school ICCs.

Although higher between-school ICCs indicate greater differences in educational opportunities across schools, higher between-school ICCs are not always the result of undesirable policies. In some cases, magnet schools may draw students with a particularly keen interest in science and mathematics. Students enrolling in these schools might have higher math test scores than students in other schools due to a combination of their interests and the focus of the school curriculum, and this could increase between-school ICCs. This would not in itself be concerning, although it would be concerning if students have differential access to these schools. Ultimately, higher between-school ICCs in a particular system (e.g., a state) may result from intentional policies or can be “emergent properties of educational systems that reflect societal values and implicit practices” (Parker et al., 2016, p. 12). More detailed evidence about between-school ICCs across states could illuminate the contexts in which variation among schools is largest. In turn, that will facilitate more accurate study design and lead to subsequent explorations of mechanisms for the variation that improve our understanding of educational opportunities.

Purpose

In this study we seek to address two primary research questions motivated by the gaps in the literature. First, *how large are within-state between-school ICCs for math and RLA test scores and how much do these ICCs vary across states, grades, years, and subject*? We report detailed descriptive statistics of within-state between-school ICCs by grade, year, subject, and state to address this question and serve as a reference for study design. We also investigate how and why our ICC estimates differ from those reported in prior studies to understand how different data and estimation methods may affect ICC estimates. Second, *what is the association between school-level ICCs and the context of schooling across states*? To better understand variation among states in the ICC estimates, we use regression analysis to examine associations between our ICC estimates and widely available school characteristics including the number of schools, average school size, percent of schools in rural and urban areas, and level of school segregation.

Data

Test Score Data

We use average math and RLA test scores in grades 3-8 for approximately 76,000 unique schools from all 50 states and the District of Columbia (DC) from 2009 to 2019 from the Stanford Education Data Archive Version 5.0 (SEDA; Reardon, Ho, et al., 2024). A full description of the construction of these school-level estimated test score means and evidence of their validity are provided in the SEDA Technical Documentation (Fahle et al., 2024). Here we summarize the most salient details about this process. Average school test score estimates in

SEDA are based on annual state accountability testing data for school years 2008-09 through 2018-19 reported in the ED*Facts* database housed by the U.S. Department of Education. The ED*Facts* assessment data files report the count of students scoring at each "proficiency" level in mathematics and RLA in each school in each state, grade, and year (e.g., U.S. Department of Education, 2020). Heteroskedastic ordered probit models are used to estimate the mean of underlying student test scores for each school based on these counts (Reardon et al., 2017; Shear & Reardon, 2021). These estimates are in standardized units relative to the student-level true score distribution (adjusting for test score measurement error) within state-grade-year-subject.[3] We describe our method of estimating ICCs from these data below.

We use data from a version of SEDA that differs from the publicly available data files in two ways. First, although SEDA uses data from the NAEP to link estimated means to a common scale across states and years, we use estimates that are standardized within state-grade-year-subject and thus do not rely on the NAEP linking. Second, the restricted-use files contain estimated test score means for individual school-grade-year-subject observations, whereas the publicly available SEDA files report school averages pooled across grades and years.

There are 51×11×6×2=6,732 possible state-grade-year-subject combinations for which we can construct ICC estimates using these data. Some state-grade-subject-years are not included

[3] As described in the SEDA technical documentation, the school mean test scores in the SEDA database are estimated adjusting for measurement error in state test scores in the following manner. First, the location of the cut scores separating proficiency levels is estimated for each state-grade-year-subject in standardized units relative to the underlying student level state test score distribution using district-level proficiency counts. These cut scores are then disattenuated to account for measurement error in the underlying student test scores by dividing the cut scores by the state test score reliability. The resulting disattenuated cut scores are used to define the metric of the school means estimated using school-level proficiency counts. In the SEDA data, these cut scores (and hence all school mean estimates) are in units standardized relative to the national NAEP distribution in each grade and subject. We reverse the national standardization via linear transformation based on the NAEP distribution in each state so that school mean estimates are standardized relative to each state's distribution (in the appropriate grade-year-subject). The estimated school test score means used in our analyses represent estimates of average student-level true scores (net of measurement error) on a metric in which student-level true scores are standardized within state-grade-year-subject.

in the SEDA data. This occurs for two primary reasons: 1) students take course-specific tests that are not comparable statewide (most commonly for eighth grade math), or 2) statewide test participation rates are below 94% and thus make estimates of the statewide test score distribution potentially inaccurate. In addition to sample exclusion criteria used in the SEDA data construction process, school test score mean estimates for specific school-grade-year-subject observations can also be missing. This occurs when: 1) the school test participation rate is less than 95%; 2) more than 40% of students take an alternate assessment; or 3) the data are insufficient for estimation (see Fahle et al., 2024).

The final sample includes roughly 3.95 million estimates of average school test scores across 6,021 total state-grade-year-subject combinations (2,942 in math and 3,079 in RLA). Table 1 reports sample coverage rates as the proportion of schools and students represented in our sample. The population count of number of schools and students is derived from publicly available NCES Common Core of Data (CCD) files, described in more detail below.[4] Due to differences in the timing of data collection between CCD and ED*Facts* data the sample coverage rates can exceed 100%. On average the ICC estimates are based on test score data for 77.0% of schools and 93.2% of students in math and 76.5% of schools and 92.6% of students in RLA. As a robustness check, we repeat some analyses after limiting the analytic sample to ICC estimates based on data from 80% or more of the schools in a state-grade-year. This reduced sample includes 3,151 total ICC estimates (1,587 in math and 1,564 in RLA). Table 1 reports sample coverage for this reduced sample. In the reduced sample the ICC estimates represent on average 88.8% of schools and 97.2% of students for math and 88.7% of schools and 96.8% of students for RLA.

[4] Available at https://nces.ed.gov/ccd/.

School Context Data

We investigate the extent to which characteristics of schools in a particular state-grade-year are associated with ICC estimates, using state-grade-year covariate data. The school context data are from SEDA school covariate data files produced based on CCD data files for the 2008-09 through 2018-19 school years following processes described by Fahle et al. (2024). For each state-grade-year the CCD data are used to calculate the total number of schools enrolling students and the average grade-level enrollment across all schools in the state. We also calculate the percent of schools listed as being in a rural locale and the percent of schools listed as being in an urban locale. These variables allow us to investigate the extent to which structural and regional variables are predictive of ICC estimates.

To explore whether school segregation is associated with variation in ICC estimates we include three measures of racial and economic segregation reported in the SEDA covariate files. We include two measures of between-school racial segregation (White-Black segregation and White-Hispanic segregation) and a between-school measure of economic segregation (based on free and reduced-price lunch eligibility). The between-school racial and economic segregation are quantified using the information theory index H (Theil, 1972). This index takes on values from 0 to 1. A value of 0 indicates no segregation and occurs when every school has a racial (or economic) composition equal to the statewide total. A value of 1 indicates complete segregation and occurs when every school enrolls students of only a single race or economic status.

All school characteristic variables are computed for every state, grade (3-8), and year (2009-2019) combination. In our regression analyses we rescale and aggregate these school characteristics in different ways. We use the natural logarithm of the number of schools and average grade level enrollment to linearize the association between these variables and the ICC estimates. For our hierarchical linear model (HLM) analyses we aggregate these variables to the state level by averaging each school characteristic variable across all 66 grade-year observations within each state. We mean-center or standardize variables to facilitate interpretation of the coefficients. Table 2 presents summary statistics and correlations among the state-level school characteristics and state-level average ICC estimates. The pattern of correlations is similar in the disaggregated data.

On average across states 39% of schools are in rural areas and 24% are in urban areas; the correlation between percent rural and urban is -0.67. In DC nearly 100% of schools are in urban areas whereas in Vermont nearly 80% of schools are rural and about 3% are urban. The average number of schools varies widely across states, from about 95 (Delaware) to over 5,000 (California). Average grade level enrollment per school ranges from about 28 (Alaska) to 163 (Georgia). The three segregation measures are positively correlated (correlations range from 0.54 to 0.75). White-Black segregation is highest on average (M=0.38) with a wide range from 0.18 (Delaware) to 0.63 (New York). White-Hispanic segregation is 0.28 on average and ranges from 0.12 (Wyoming) to 0.56 (DC). Economic segregation is 0.18 on average and has a smaller range from 0.05 (West Virginia) to 0.33 (Connecticut). The three segregation measures are negatively correlated with the percent of rural schools and positively correlated with the percent of urban schools. The average ICC estimates in math and RLA are highly correlated across states (*r*=0.93). Among the covariates, economic segregation is most highly correlated with the average ICC estimates in both subjects. All three segregation measures are positively correlated with ICC estimates, as are the number of schools, average school enrollment, and percent of urban schools. Only the percent of rural schools is negatively correlated with ICC estimates.

Methods

ICC Estimates

The ICC is most commonly estimated using student-level test score data within an HLM framework (Raudenbush & Bryk, 2002). With student-level data, a random-intercept model of the form

$$y_{is} = \gamma + \beta_s + \epsilon_{is}$$

$$\beta_s \sim N(0, \sigma_B^2), \epsilon_{is} \sim N(0, \sigma_W^2)$$

would be estimated, where y_{is} is the test score for student i in school s, the β_s are normally distributed random school intercepts, and the ϵ_{is} are normally distributed student-level errors. The ICC, ρ, would be estimated as the proportion of total variance that is between schools using the expression

$$\hat{\rho} = \frac{\hat{\sigma}_B^2}{\hat{\sigma}_B^2 + \hat{\sigma}_W^2}.$$

The ICC can also be defined within an analysis of variance (ANOVA) framework as the proportion of total variance between schools taking into account school sample sizes and potentially unequal within-school variances (for a discussion see Fahle & Reardon, 2018). We use the approach described below, based on the HLM framework, for consistency with prior studies and because we do not have precisely estimated within-school variances.

Because no student-level data are available from SEDA, we estimate the ICC for each state-grade-year-subject based on SEDA school mean test score estimates using a precision-weighted random effects model, also referred to as a V-known model (Raudenbush & Bryk, 2002). We estimate the following model separately within each state, grade, year, and subject:

$$\hat{\mu}_s = \beta + u_s + \epsilon_s$$

$$u_s \sim N(0, \tau^2), \epsilon_s \sim N(0, \hat{\omega}_s^2),$$

where $\hat{\mu}_s$ is the estimated mean test score in school s and $\widehat{\omega}_s^2$ is the estimated sampling variance of the school mean test score, each from the SEDA data. We treat the value $\widehat{\omega}_s^2$ (the variance of ϵ_s) as known. The school means, $\hat{\mu}_s$, are reported on a scale that is standardized relative to the student-level distribution of true (error-free) scores within each state-grade-year-subject, as described above. The total variance (between-schools plus within-schools) of scores at the student-level on this scale is equal to 1. Thus, τ^2 represents the true variance in school mean test scores after accounting for sampling error in the estimated means, which is the same target quantity estimated by $\hat{\sigma}_B^2$ when student-level test scores are standardized. We estimate τ^2 via maximum likelihood using the –`metareg`– package for Stata (Harbord & Higgins, 2008). We use $\hat{\tau}^2$ as an estimate of ρ, the ICC representing the proportion of variance that is between school variance from a two-level unconditional model with students nested in schools (Raudenbush & Bryk, 2002).

These ICC estimates differ from estimates derived from an unconditional two-level HLM model fit to student-level test score data, which is relevant when using the ICC estimates reported here or comparing them to prior (and future) research. First, because we rely on school mean test score estimates in a standardized metric from SEDA, our estimate $\hat{\tau}^2$ is most accurately described as the standardized between-school variance. We use this as a proxy for the between-school ICC and refer to these estimates as ICCs in the remainder of the paper, but it is not a direct estimate of the ICC that would be produced by a two-level HLM model. Second, because heteroskedastic ordered probit models are used to construct the mean test score estimates in SEDA, the average school test scores are not reported on the original state test score scale but instead on a respectively normal metric (Reardon & Ho, 2015). The true ICC in this metric is not necessarily identical to the true ICC based on the original test score scale. Third, the

SEDA data construction process estimates school means in a standardized metric after correcting for test score measurement error. When the ICC from a two-level HLM model is estimated based on observed test scores containing measurement error, the total variance will be inflated while the estimate of $\hat{\sigma}_B^2$ is unbiased and hence the estimated ICC is attenuated (Cox & Kelcey, 2019). The ICC among observed scores, ρ_{obs}, will be equal to $\rho_{true} * rel(x)$, where ρ_{true} is the ICC among error-free scores and $rel(x)$ is the reliability of the observed scores. The ICC estimates we report can be interpreted as an indicator of the ICC among error-free scores that have measurement error variance removed. While this may limit comparability to previously reported ICCs, a benefit is that these ICC estimates are not confounded by differences in test score reliability across states, grades, and years.

We compare our ICC estimates to two other sources of school-level ICC estimates from small groups of states that overlap with a subset of the states, grades, and years included in our sample (Hedges & Hedberg, 2014; Westine et al., 2013) and to estimates of national-level between school ICC estimates (Hedges & Hedberg, 2007). The comparison to prior studies estimating between-school ICCs for the same states, grades, and years provides a validation check of our methodology and sheds light on the reasons that ICC estimates can vary across studies that nominally estimate the same ICCs using the same data source.

Statistical Analyses of ICC Estimates and School Context

We investigate the extent to which the context of schooling, indicated by school characteristics, is associated with ICC estimates using two different approaches. First, we use a series of ordinary least squares (OLS) regression models across the full, disaggregated dataset pooling across subjects to characterize the association between school characteristics and ICC estimates. The OLS models may be useful to those planning intervention studies. If ICC estimates are associated with school characteristics that study designers know ahead of time, it allows them to more accurately predict the relevant ICC for the context of their study. For these OLS models we center each variable relative to the mean in the full sample.

Second, to better understand how school characteristics are associated with variation across states and how ICCs have changed over time, we use a series of two-level HLM regression models to model variation in ICC estimates within and between states. These models parallel those used by Fahle and Reardon (2018) to study variation in between-district ICCs across states. Specifically, we use a two-level HLM specification with ICC estimates at level 1 (varying across grades and years) and states at level 2. We estimate these models separately for each subject. The first model we specify has the following form at level 1:

$$y_{sgy} = \beta_{0s} + \beta_{1s}\left(grade_{sgy} - \overline{grade_s}\right) + \beta_{2s}\left(year_{sgy} - \overline{year_s}\right) + \delta_{ys} + e_{sgy}$$

where y_{sgy} is the estimated ICC in state $s,$ grade g, and year y and the δ_{ys} are year dummy variables, with 2009 (the earliest year) as the omitted category. We mean-center the $grade$, $year$, and year dummy variables within state-subject so that these variables characterize changes in ICC estimates within states. The level 2 model (state level) has the form

$$\beta_{0s} = \gamma_{00} + u_{0s}$$

$$\beta_{1s} = \gamma_{10} + u_{1s}$$

$$\beta_{2s} = u_{2s}$$

$$(u_{0s}, u_{1s}, u_{2s}) \sim N\left(\mathbf{0}, \boldsymbol{\tau} = \begin{bmatrix} \tau_{00} & & \\ \tau_{01} & \tau_{10} & \\ \tau_{02} & \tau_{12} & \tau_{20} \end{bmatrix}\right), e_{sgy} \sim N(0, \sigma^2)$$

where u_{0s}, u_{1s}, and u_{2s} are normally distributed random effects. The equation for β_{2s} has no fixed component, constraining the linear year trend to have a mean of 0 across states. This allows us to include year dummies for all years 2010-2019 and a state-specific linear deviation from these nonparametric year means. Preliminary analyses confirmed that the variance of the state grade and year slopes (τ_{10} and τ_{20}, respectively) were statistically different from 0. This model is used to quantify the variability in average ICC estimates across states (represented by τ_{00}), net of state-specific linear grade trends and differences across years.

In Models 2-4 we add state level covariates as predictors of the state intercepts. The state level covariates are constructed by averaging across all grades and years within each state and standardizing the average value relative to the distribution of averages across states. The resulting coefficients can be interpreted as the predicted difference in a state's average ICC estimate associated with a state-level standard deviation change in the covariate. In Model 2 we include school structure variables representing the natural log of number of schools, natural log of average grade level enrollment, and percent of rural schools. In Model 3 we include two segregation measures (Free Lunch segregation and White-Black segregation).[5] In Model 4 we include both the school structure variables and segregation measures. These models quantify the extent to which the state-level school characteristics can explain cross-state variation in average ICC estimates. All descriptive analyses were carried out in the R computing environment (R Core Team, 2024) and the HLM models were estimated via maximum likelihood using the `lme4` package in R (Bates et al., 2015). A data file containing all state-grade-year-subject ICC estimates is available at the link provided in the Data Availability Statement.

Results

Table 3 presents summary statistics for the ICC estimates, averaged by subject, for the full set of estimates and the reduced set of estimates based on at least 80% of schools. In the full

[5] In preliminary analyses we included the percent of urban schools variable and White-Hispanic segregation. However, these variables explained little additional variance beyond the other variables and the coefficients were difficult to interpret due to collinearity. We thus exclude them from the HLM models for parsimony.

sample, average ICC estimates are higher in math (M=0.194, SD=0.056) relative to RLA (M=0.168, SD=0.054) and have a larger range (0.056 to 0.399 in math versus 0.049 to 0.338 in RLA). Estimates in the reduced sample are similar and also relatively larger in math (M=0.204, SD=0.055) than in RLA (M=0.177, SD=0.052). All subsequent results are based on the full sample of ICC estimates. The Supplementary Materials include additional descriptive tables of ICC estimates and analyses using the reduced sample. Results based on the reduced sample were substantively similar to those using the full sample.

Table 4 compares between-school ICC estimates from three prior studies for which a direct comparison is informative: (a) Hedges and Hedberg (2007); (b) Hedges and Hedberg (2014); and (c) Westine et al. (2013). The first two columns compare our ICC estimates (averaged across states) to the national between-school estimates reported by Hedges & Hedberg (2007). Our average grade-level estimates are smaller than the estimates reported by Hedges and Hedberg (2007) for all grades and subjects except grade 8 math. The differences are larger in elementary grades and smaller for middle school grades. Potential reasons for these differences include that the Hedges and Hedberg (2007) estimates are sensitive to between-state differences, sampling error due to the use of national probability samples, changes to ICCs over time, and the specific tests used. These differences, and the variability across states, indicate that use of national estimates may not be appropriate for planning studies in a single state using more recent state assessments.

The next four columns in Table 4 compare our state-level estimates to estimates in Hedges and Hedberg (2014). Our data overlap for nine states: seven states in 2009-10 (AR, AZ, KS, KY, MA, NC, WI), one state in 2012-13 (LA), and one state in 2011-12 (WV). Only eight states are included for grade 8 math comparisons because grade 8 math in AR is not included in our sample. We report grade level averages, correlations, mean differences, and root mean squared differences (RMSD) between the estimates. Our estimates are highly correlated with those reported in Hedges and Hedberg, with an average correlation of 0.946 and average RMSD of 0.025. The average difference is 0.006, suggesting the direction of differences is not systematic. The final two columns of Table 4 show the comparison to Westine et al. (2013). We compare the Westine et al. (2013) estimates to our grade 5 and 8 ICC estimates from Texas averaged over 2008-09 to 2010-2011. The four pairs of estimates are extremely close, except for 8th grade RLA, for which the estimate based on our data (0.127) is higher than that reported by Westine et al. (0.099).

While the state-specific studies nominally estimate the same ICCs using the same data that we use (except for averaging over different years in Texas), there are at least four reasons our ICC estimates might differ. First, our estimates represent ICCs on a respectively normal metric rather than the original test score scales in each state. Second, our ICCs are based on aggregate SEDA data rather than variance components obtained via two-level HLM models applied to student-level data, which were used in both prior studies. Third, our ICC estimates are disattenuated to account for measurement error. Fourth, each study makes different sample restrictions and data cleaning decisions, so that the analytic samples of students used to estimate ICCs differs. Hedges and Hedberg (2014), for example, exclude students with cognitive disabilities and students attending charter schools from their analyses. Westine et al., (2013) removed approximately 16% of students in each grade and year due to data cleaning and an additional 10-20% due to masking of small samples. Our sample includes about 5% fewer schools in each state and 15% more students relative to the Hedges and Hedberg samples and includes approximately the same number of schools but between 30% and 70% more students relative to the Westine et al. samples.

We turn now to our investigation of which school characteristics are correlated with ICC estimates. Table 5 presents OLS regression results. The school context variables together explain approximately 58% of the variance in ICC estimates. In other words, there is meaningful variation in the between-school ICCs across states, grades, and years and a substantial proportion of this variation can be explained by structural differences in the school system across states and grades. The ICCs are about 0.026 units smaller in RLA relative to math, on average, and this difference remains constant when adjusting for other covariates. The strongest individual predictor of ICCs (measured by R^2) is between-school economic segregation ($R^2 = 0.468$) followed by the percent of rural schools ($R^2 = 0.349$). On average, ICC estimates are larger when there are more schools in urban settings and smaller when there are more schools in rural settings. Although the estimated ICCs differed little across grades in the overall sample (see Column 2), the ICCs in higher grades tended to be larger on average when adjusting for the additional covariates (see Column 10). This suggests differences in ICCs across grades are partly due to school structures and characteristics.

The sign of the coefficients for percent urban schools and White-Hispanic segregation reverse when including all covariates, due to collinearity with other variables. We include these variables in the OLS models to facilitate the most accurate predictions of ICCs but exclude them from the HLM models below for parsimony and clearer interpretations. The percent of urban schools and White-Hispanic segregation increase R^2 by less than 0.001 and by 0.007, respectively, beyond other variables. In contrast, the percent of rural schools, White-Black segregation, and free lunch segregation increase R^2 by 0.043, 0.044, and 0.087, respectively, beyond other variables.

The HLM models in Table 6 indicate that most of the variance in ICC estimates is between states – conditional on grade and year approximately 85% of the unexplained variance ICC estimates is between states in each subject. The within-state standard deviations of ICC estimates are 0.021 and 0.019 in math and RLA, respectively, while the between-state standard deviations are 0.049 and 0.048. Figure 1 displays a map of Empirical Bayes (EB) estimates of average ICCs from Model 1 for all 50 states and DC. The highest average ICC estimates were in the Northeast and Midwest; West Virginia had the lowest average ICC estimate in both subjects. Average ICC estimates decline slightly across grades in RLA (-0.002 points per grade), but not in math. The structural variables explain about 45% and 54% of the between-state variance of ICC estimates. States with more schools, lower average grade level enrollment, and smaller shares of rural schools have higher ICCs on average in both subjects. The segregation measures explain a greater proportion of the between-state variance in ICC estimates – about 66% and 70% in math and RLA, respectively. Between-school ICCs are higher in states with higher levels of between school White-Black racial segregation and economic segregation, on average.

When combined, the school structure variables and segregation measures explain approximately 70% and 78% of the between-state variance in average ICCs in math and RLA, respectively. Controlling for school structures and locale, states with higher levels of racial and economic segregation have higher between-school ICC estimates, on average. In math, coefficients for racial and economic segregation are similar in magnitude, while in RLA the coefficient for economic segregation is approximately twice as large as the coefficient for racial segregation. In math the only school structure variable that remains a statistically significant predictor is grade level enrollment (ICCs are lower in states with larger schools). In RLA the same association with school size is observed, while ICCs are also lower in states with more rural schools. Taken together, the level of segregation is more predictive of between-school ICCs than are features of school structures.[6] Figure 2 plots state ICC EB estimates from Model 1 versus the state-level covariates to show these relationships visually.

The HLM models also quantify changes in ICC estimates over time. The year coefficients in Table 6 estimate the average within-state difference of ICCs for each year relative to 2009 in the middle grade for each state (approximately midway through 5th grade), net of other variables in the model. Beginning in 2015, these year coefficients are consistently positive and statistically significant, and slightly larger in math than in RLA on average. To investigate these trends further, we fit an additional HLM model for each subject and grade. These models were equivalent to Model 1 but exclude the grade trends (because the models were fit separately for each grade). Figure 3 presents estimates of the year indicator variables from these models (which continue to represent average within-state differences relative to 2009), with error bars denoting 99% confidence intervals. Average ICCs increase from 2009 to 2019 in grades 3-6 but not in grades 7 and 8. In math, there were consistent increases in grades 3 and 4 during 2010-2014 that leveled off in 2015-2019. This pattern was less pronounced in grades 5-6 math and in RLA. Table 7 presents the ICC estimates averaged by subject, grade, and year. The marginal row and column means report averages across cells, unweighted by the number of state estimates in each grade-year cell.

[6] We also fit HLM models that included school structure and segregation measures as variables that vary within-states. However, these variables explained less than 5% of the residual within-state variance suggesting little systematic within-state associations.

Comparisons across grades and years should be interpreted with some caution due to changing sample sizes, although the impact of this missing data is limited because our model estimates within-state changes. For example, in 2009 there are 49 and 44 state math ICC estimates in grades 3 and 8, respectively, while in 2019 there are 46 and 36 estimates in grades 3 and 8, respectively. When limiting to a balanced sample of states across years within each grade a similar pattern was observed.

Discussion

This paper makes both methodological and substantive contributions. Methodologically, we provide the most extensive evidence to date about within-state between-school ICCs of standardized test scores using 11 years of U.S. population data. The wide variation in ICCs (ranging from 0.049 to 0.399) underscores prior findings that single rule of thumb ICC values may not be adequate for planning studies in all contexts, leading to studies being grossly over- or under-powered. For example, to detect as significant an effect size of 0.20 in a two-level CRT with average cluster size of 86 (the average school enrollment in our sample) and no covariates, we would need 50 clusters if the outcome ICC were 0.05 but 322 clusters if the outcome ICC were 0.39 (Dong & Maynard, 2013). The detailed tables provided in the Supplementary Materials and OLS regression results could be used to more accurately predict between-school ICCs, informing study design in other contexts as well.

Substantively, we found significant variability in between-school ICC estimates, with most of the variation existing between states. The variation in ICCs among states is systematically linked to the context of schooling, as indicated by the high proportion of variance in the ICCs explained by a relatively small number of variables. The strong association between ICCs and segregation is consistent with the hypothesis that between-school segregation is connected to unequal access to educational opportunities. Although we cannot differentiate the extent to which ICCs are driven by differences in school-related factors versus selection of students into schools or other out of school factors, the results clearly show states with more segregated schools also have greater variability of educational outcomes across schools.

Our findings mirror those reported based on between-district ICC analyses, including the strong association between economic segregation and ICCs (Fahle & Reardon, 2018). Relative to between-district ICCs, between-school ICCs were larger on average, had greater variability within states, and had similar variability between states. The proportion of variance in between-school ICCs across states explained by school context was slightly smaller than the proportion of variance explained in between-district ICCs. Consistent with prior empirical evidence about between-school ICCs, our ICC estimates were higher for math than for RLA tests, a pattern that was true across states, grades, and years. In contrast to some prior empirical data, we did not find a consistent pattern in changes of between-school ICCs across grade levels, although ICCs in RLA declined on average across grades within states. School structures (indicated by the number of schools and school enrollments) were inconsistently associated with between-school ICCs depending upon the additional factors included in the models, although states with more schools in rural areas had lower average ICCs.

The increase in ICCs over time has not previously been reported, although it is consistent with increases in between-district ICCs reported by Fahle and Reardon (2018). Whereas Fahle and Reardon report a roughly linear increase in between-district ICCs across cohorts of students observed between 2009 and 2015 and averaged across grade levels, we estimated changes across a longer time period and focused on comparisons across years within grades rather than cohorts (noting that within each grade, cohort and year trends are equivalent). In both subjects, we observed increases in grades 3 to 6 of about 10-18 percent, primarily after 2015, with little change in grades 7 and 8. These changes seem more consistent with changes to instruction and assessment rather than changes in segregation or school structures. Elementary school structures did not change meaningfully in 2015, and from 2009 to 2019 in our data segregation changed less than ICCs and changes were not differential across grade levels. Within grades, White-Black segregation increased by about 6-9 percent, Free Lunch segregation increased by about 1-2 percent, and White-Hispanic segregation declined by about 5-8 percent.

Many states adopted the Common Core State Standards (CCSS) and implemented new tests aligned with these standards beginning in 2014 and 2015. The types of changes brought about by CCSS could have differed across grade levels. If increases in ICCs are related to CCSS implementation, it raises the question as to whether the increased ICCs are more a reflection of growing disparities in student outcomes or of the constructs assessed by new tests. During the decade 2009-2019 the gap between the 90th percentile and 10th percentile (90/10 gap) of the NAEP score distribution increased at grades 4 and 8 in both math and reading (Wilburn et al., 2019). Although not directly tied to ICCs, this growing divergence on a test that remained constant (and is not directly aligned to any single state assessment) is more consistent with the hypothesis that learning outcomes have become more unequal across schools. On the other hand, the increasing 90/10 gap on NAEP was observed in both subjects at grades 4 and 8, which is inconsistent with the differential changes we observed across grade levels. It is possible that changes to instruction and assessments brought about by the CCSS affected schools differentially and thus increased ICCs, but the impact of CCSS on student achievement continues to be debated (e.g., Briggs et al., 2024). We cannot disentangle the various causes of changes to ICCs and this remains an important avenue for further study.

There are relevant limitations of our results that should also be noted. The SEDA data do not allow us to directly estimate the between-school ICC that could be estimated with student-level data. The consistency of our reported estimates with previously reported estimates based on student-level data and research validating the statistical methods used to construct the SEDA data (Fahle et al., 2024 and citations therein) support use of these estimates as proxies for between-school ICCs. However, those using the ICC estimates reported here should keep in mind that the methods used to estimate ICCs reported here make different assumptions than prior research reporting between-school ICCs. To the extent there were differences between our estimates and previously reported values, our comparisons emphasize the importance of attending to these data processing and analytic choices. In the context of ICC estimates, even when researchers may have access to an ICC estimate that is directly relevant to their study context—in terms of the population, context, and test—the comparability of data analytic choices must also be considered. For example, measurement error will attenuate ICCs estimated using a two-level HLM model, but this was rarely discussed in prior studies. Missing data may limit generalizability of our results, although the findings we report are robust to different sample restrictions suggesting the patterns reported are not primarily driven by systematic missingness. Finally, when student-level test score data are available, estimated ICCs based on these data should be preferred. If researchers have access to statewide student-level data in the context where they will conduct a study, we recommend calculating an ICC estimated based on those data rather than using the ICCs reported here.

Conclusion

We document substantial variability of within-state between-school ICCs across states from 2009 to 2019. The average magnitudes of these ICCs (0.194 in math and 0.168 in RLA) and their ranges are consistent with between-school ICCs reported in prior literature, but the scope of the estimates presented here provides new insights into the variability of between-school ICCs across states and changes over time. Investigating the apparent increase in between-school ICCs beginning in 2015 is an important avenue for future work. The between-school ICCs are highly correlated with the level of school economic and racial segregation in a state. These patterns are consistent with prior research documenting the association between school segregation and disparities in educational outcomes (e.g., Reardon, Weathers, et al., 2024) and suggest policies aimed at reducing school segregation may be an important avenue to improve the equality of educational opportunities across schools.

We conclude by noting another important topic for future research. The SEDA data do not include school level estimates for years after the COVID-19 pandemic that began in 2020. Investigating the extent to which ICCs have changed since 2020 will help identify how educational opportunities have shifted in recent years and inform the design of future studies. Given changes to federal data collection of state assessment results, studying this variation at the national level will require alternative data sources. One possibility is that states and test vendors responsible for publicly reporting state assessment results could report between-school ICCs as part of public dashboards or as part of public technical reports. This would require relatively little additional analysis and would provide a valuable resource to a wide range of stakeholders. Our results provide a detailed baseline reference point for such future studies and will be valuable for a broad range of educational statisticians, ranging from methodologists designing studies to educational researchers or state personnel focused on understanding equity among schools.

References

Angrist, J. D., Hull, P. D., Pathak, P. A., & Walters, C. R. (2017). Leveraging lotteries for school value-added: Testing and estimation. *The Quarterly Journal of Economics*, *132*(2), 871–919. https://doi.org/10.1093/qje/qjx001

Bates, D., Mächler, M., Bolker, B. M., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1). https://doi.org/10.18637/jss.v067.i01

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

Brandon, P. R., Harrison, G. M., & Lawton, B. E. (2013). SAS code for calculating intraclass correlation coefficients and effect size benchmarks for site-randomized education experiments. *American Journal of Evaluation*, *34*(1), 85–90. https://doi.org/10.1177/1098214012466453

Briggs, D. C., Shepard, L., & Buchbinder, N. (2024). *What can NAEP mathematics subscales and subscale weights tell us about Common Core effects?* (NAEP Validity Studies Panel).

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. *American Economic Review*, *104*(9), 2593–2632. https://doi.org/10.1257/aer.104.9.2593

Cox, K., & Kelcey, B. (2019). Optimal design of cluster- and multisite-randomized studies using fallible outcome measures. *Evaluation Review*, *43*(3–4), 189–225. https://doi.org/10.1177/0193841X19870878

Dogan, E., Walton, E., Reardon, S. F., Broer, M., Bai, Y., & Zheng, X. (2024, June 24). *Appraising the state of education equity in the U.S.: A conceptual framework and quantitative illustrations*. CCSSO 2024 National Conference on Student Assessment, Seattle, WA.

Dong, N., & Maynard, R. (2013). *PowerUp!*: A tool for calculating minimum detectable effect sizes and minimum required sample sizes for experimental and quasi-experimental design studies. *Journal of Research on Educational Effectiveness*, *6*(1), 24–67. https://doi.org/10.1080/19345747.2012.673143

Dong, N., Reinke, W. M., Herman, K. C., Bradshaw, C. P., & Murray, D. W. (2016). Meaningful effect sizes, intraclass correlations, and proportions of variance explained by covariates for planning two- and three-level cluster randomized trials of social and behavioral outcomes. *Evaluation Review*, *40*(4), 334–377. https://doi.org/10.1177/0193841X16671283

Fahle, E. M., & Reardon, S. F. (2018). How much do test scores vary among school districts? New estimates using population data, 2009–2015. *Educational Researcher*, *47*(4), 221–234. https://doi.org/10.3102/0013189X18759524

Fahle, E. M., Saliba, J., Kalogrides, D., Shear, B. R., Reardon, S. F., & Ho, A. D. (2024). *Stanford Education Data Archive: Technical documentation (Version 5.0)*. https://purl.stanford.edu/cs829jn7849

Gray, H. L., Burgermaster, M., Tipton, E., Contento, I. R., Koch, P. A., & Di Noia, J. (2016). Intraclass correlation coefficients for obesity indicators and energy balance–related behaviors among New York City public elementary schools. *Health Education & Behavior*, *43*(2), 172–181. https://doi.org/10.1177/1090198115598987

Harbord, R. M., & Higgins, J. P. T. (2008). Meta-regression in Stata. *Stata Journal*, *8*(4), 493–519.

Hedberg, E. C., & Hedges, L. V. (2014). Reference values of within-district intraclass correlations of academic achievement by district characteristics: Results from a meta-analysis of district-specific values. *Evaluation Review*, *38*(6), 546–582. https://doi.org/10.1177/0193841X14554212

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

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

Juras, R. (2016). Estimates of intraclass correlation coefficients and other design parameters for studies of school-based nutritional interventions. *Evaluation Review*, *40*(4), 314–333. https://doi.org/10.1177/0193841X16675223

Koedel, C., Mihaly, K., & Rockoff, J. E. (2015). Value-added modeling: A review. *Economics of Education Review*, *47*, 180–195. https://doi.org/10.1016/j.econedurev.2015.01.006

Parker, P. D., Jerrim, J., Schoon, I., & Marsh, H. W. (2016). A multination study of socioeconomic inequality in expectations for progression to higher education: The role of between-school tracking and ability stratification. *American Educational Research Journal*, *53*(1), 6–32. https://doi.org/10.3102/0002831215621786

R Core Team. (2024). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Sage Publications, Inc.

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 changes* (pp. 91–116). 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*, *40*(2), 158–189. https://doi.org/10.3102/1076998615570944

Reardon, S. F., Ho, A. D., Shear, B. R., Fahle, E. M., Kalogrides, D., & Saliba, J. (2024). *Stanford Education Data Archive* (Version 5.0) [Dataset]. https://purl.stanford.edu/cs829jn7849

Reardon, S. F., Robinson-Cimpian, J. P., & Weathers, E. S. (2015). Patterns and trends in racial/ethnic and socioeconomic academic achievement gaps. In H. F. Ladd & M. E. Goertz (Eds.), *Handbook of research in education finance and policy* (2nd ed., pp. 491–509). Routledge.

Reardon, S. F., Shear, B. R., Castellano, K. E., & Ho, A. D. (2017). Using heteroskedastic ordered probit models to recover moments of continuous test score distributions from coarsened data. *Journal of Educational and Behavioral Statistics*, *42*(1), 3–45. https://doi.org/10.3102/1076998616666279

Reardon, S. F., Weathers, E. S., Fahle, E. M., Jang, H., & Kalogrides, D. (2024). Is separate still unequal? New evidence on school segregation and racial academic achievement gaps. *American Sociological Review*, *89*(6), 971–1010. https://doi.org/10.1177/00031224241297263

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. (2021). Using pooled heteroskedastic ordered probit models to improve small-sample estimates of latent test score distributions. *Journal of Educational and Behavioral Statistics*, *46*(1), 3–33. https://doi.org/10.3102/1076998620922919

Shen, Z., Curran, F. C., You, Y., Splett, J. W., & Zhang, H. (2023). Intraclass correlations for evaluating the effects of teacher empowerment programs on student educational outcomes. *Educational Evaluation and Policy Analysis*, *45*(1), 134–156. https://doi.org/10.3102/01623737221111400

Theil, H. (1972). *Statistical decomposition analysis*. North-Holland.

U.S. Department of Education. (2020). *State assessments in reading/language arts and mathematics: School year 2018-19 EDFacts data documentation*. U.S. Department of Education. http://www.ed.gov/edfacts

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

Wilburn, G., Cramer, B., & Walton, E. (2019, October 30). *The great divergence: Growing disparities between the nation's highest and lowest achievers in NAEP mathematics and reading between 2009 and 2019*. https://nces.ed.gov/nationsreportcard/blog/mathematics_reading_2019.aspx

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

Tables and Figures

Table 1

Sample Coverage Rates by Subject

Sample	Subject	Variable	N	Mean	SD	Median	Min	Max
Full Sample	Math	% Schools	2942	0.770	0.163	0.811	0.184	0.987
		% Students	2942	0.932	0.078	0.961	0.471	1.052
	RLA	% Schools	3079	0.765	0.160	0.802	0.175	0.988
		% Students	3079	0.926	0.081	0.956	0.478	1.082
80% Sample	Math	% Schools	1587	0.888	0.050	0.889	0.800	0.987
		% Students	1587	0.972	0.030	0.979	0.835	1.052
	RLA	% Schools	1564	0.887	0.050	0.889	0.800	0.988
		% Students	1564	0.968	0.034	0.976	0.816	1.082

Note. ICC=intraclass correlation coefficient. RLA= Reading/Language Arts. Observations are state-subject-grade-years. Coverage rates are calculated as the number of test scores used to calculate the ICC for the state-subject-grade-year from SEDA divided by the population student counts from CCD for that state-subject-grade-year. Sample coverage rates in some state-subject-grade-years can exceed 1.0 due to differences in two data sources.

Table 2

Descriptive Statistics and Correlations Among School Characteristics and ICC Estimates

Row	Variable	M	SD	Min	Max	1	2	3	4	5	6	7	8	9	10	11
1	ICC Math	0.20	0.05	0.09	0.31	1.00										
2	ICC RLA	0.17	0.05	0.08	0.28	0.94	1.00									
3	N Schools	831.1	853.0	95.3	5021.2	0.29	0.29	1.00								
4	Avg. Enrollment	86.1	30.1	27.7	162.9	0.18	0.17	0.28	1.00							
5	ln(Schools)	6.33	0.86	4.52	8.48	0.33	0.29	0.84	0.31	1.00						
6	ln(Avg. Enrollment)	4.35	0.39	3.32	5.03	0.23	0.19	0.32	0.96	0.37	1.00					
7	% Rural	0.39	0.20	0.00	0.80	-0.60	-0.64	-0.36	-0.60	-0.30	-0.67	1.00				
8	% Urban	0.24	0.15	0.03	1.00	0.52	0.50	0.27	0.19	0.14	0.24	-0.67	1.00			
9	White-Black Segregation	0.38	0.13	0.18	0.63	0.65	0.57	0.40	0.22	0.52	0.30	-0.43	0.47	1.00		
10	White-Hispanic Segregation	0.28	0.11	0.12	0.56	0.65	0.67	0.40	0.30	0.36	0.35	-0.67	0.64	0.75	1.00	
11	Free Lunch Segregation	0.18	0.06	0.05	0.33	0.77	0.83	0.35	0.29	0.39	0.32	-0.64	0.50	0.54	0.75	1.00

Note. ICC=intraclass correlation coefficient. RLA=reading/language arts. ln=natural log. Observations are states plus DC (N=51). Variables are averages across grades and years within states. School structure covariates are calculated for all grades and years, not only grades and years with ICC estimates.

Table 3

Summary Statistics of ICC Estimates by Subject

Sample	Subject	N	Mn	SD	Min	Max
Full Sample	Math	2942	0.194	0.056	0.055	0.399
	RLA	3079	0.168	0.054	0.049	0.338
80% Sample	Math	1587	0.204	0.055	0.056	0.397
	RLA	1564	0.177	0.052	0.055	0.320

Note. ICC=intraclass correlation coefficient. RLA= Reading/Language Arts. Observations are state-subject-grade-years.

Table 4

Comparison to Prior Reported ICC Estimates

		Hedges & Hedberg 2007		Hedges & Hedberg 2014				Westine et al. 2013	
Subject	**Grade**	**HH07**	**SEDA**	**HH14**	**SEDA**	**r**	**RMSD**	**WST**	**SEDA**
Math	3	0.241	0.190	0.165	0.179	0.967	0.020		
	4	0.232	0.198	0.169	0.178	0.987	0.013		
	5	0.216	0.202	0.177	0.192	0.982	0.019	0.168	0.160
	6	0.264	0.195	0.175	0.185	0.955	0.021		
	7	0.191	0.184	0.188	0.184	0.946	0.025		
	8	0.185	0.194	0.203	0.187	0.912	0.047	0.163	0.167
RLA	3	0.271	0.172	0.150	0.160	0.918	0.020		
	4	0.242	0.174	0.156	0.173	0.956	0.024		
	5	0.263	0.174	0.157	0.172	0.923	0.025	0.156	0.156
	6	0.260	0.165	0.152	0.160	0.945	0.020		
	7	0.174	0.162	0.164	0.164	0.959	0.026		
	8	0.197	0.161	0.171	0.164	0.896	0.043	0.099	0.127

Note. ICC=intraclass correlation coefficient. RLA=Reading/Language Arts. RMSD = root mean squared difference. r = Pearson correlation. The SEDA comparison estimates for HH07 are averaged across all available SEDA estimates. SEDA comparison estimates for HH14 are averaged across 9 states (AR, AZ, KS, KY, LA, MA, NC, WI, WV; except grade 8 math, missing AR) for the years described in text. SEDA comparison estimates for WST are from TX averaged across 2009-2011.

Table 5

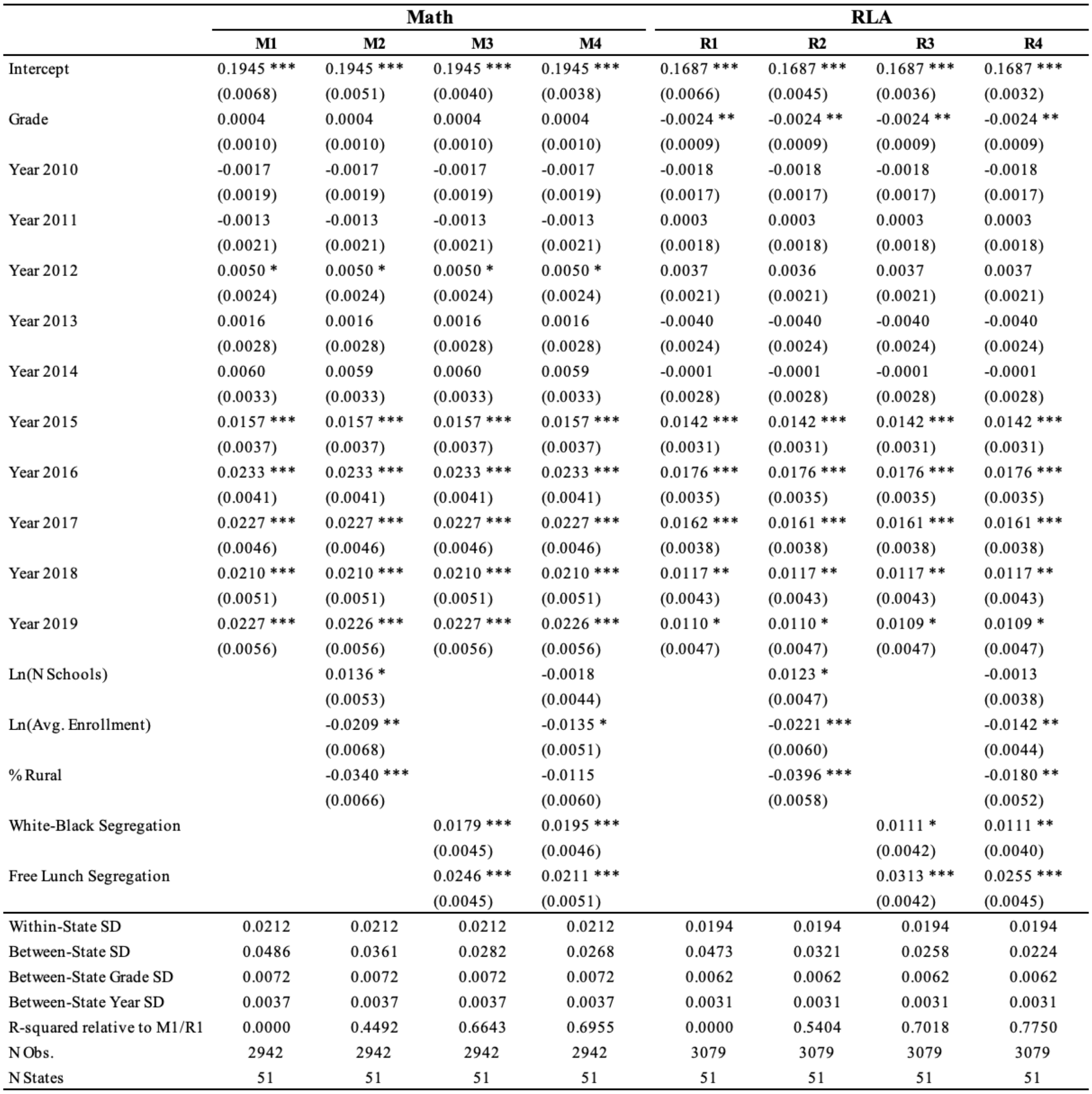
OLS Regression Model Estimates

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Intercept	0.194 ***	0.194 ***	0.194 ***	0.194 ***	0.194 ***	0.194 ***	0.194 ***	0.194 ***	0.194 ***	0.176 ***
	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
RLA	-0.026 ***	-0.026 ***	-0.026 ***	-0.026 ***	-0.027 ***	-0.026 ***	-0.026 ***	-0.026 ***	-0.026 ***	-0.026 ***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Grade 4		0.005 *								0.006 ***
		(0.002)								(0.002)
Grade 5		0.007 **								0.011 ***
		(0.002)								(0.002)
Grade 6		0.000								0.028 ***
		(0.002)								(0.002)
Grade 7		-0.008 **								0.032 ***
		(0.002)								(0.002)
Grade 8		-0.004								0.036 ***
		(0.002)								(0.002)
Ln(N Schools)			0.017 ***							-0.005 ***
			(0.001)							(0.001)
Ln(Avg. Enrollment)				0.015 ***						-0.030 ***
				(0.002)						(0.002)
% Rural					-0.149 ***					-0.113 ***
					(0.003)					(0.005)
% Urban						0.171 ***				-0.013 **
						(0.004)				(0.005)
White-Black segregation							0.221 ***			0.147 ***
							(0.005)			(0.006)
White-Hispanic segregation								0.284 ***		-0.087 ***
								(0.005)		(0.008)
Free Lunch Segregation									0.540 ***	0.385 ***
									(0.008)	(0.011)
N	6021	6021	6021	6021	6021	6021	6021	6021	6021	6021
R^2	0.053	0.061	0.121	0.067	0.349	0.245	0.312	0.346	0.468	0.578

Note. ****p* < 0.001; ***p* < 0.01; **p* < 0.05. RLA=Reading/Language Arts. Ln=natural log. Observations are state-grade-subject-years. All covariates are grand mean centered except subject and grade fixed effects.

Table 6

HLM Regression Model Estimates



	Math				RLA			
	M1	M2	M3	M4	R1	R2	R3	R4
Intercept	0.1945 ***	0.1945 ***	0.1945 ***	0.1945 ***	0.1687 ***	0.1687 ***	0.1687 ***	0.1687 ***
	(0.0068)	(0.0051)	(0.0040)	(0.0038)	(0.0066)	(0.0045)	(0.0036)	(0.0032)
Grade	0.0004	0.0004	0.0004	0.0004	-0.0024 **	-0.0024 **	-0.0024 **	-0.0024 **
	(0.0010)	(0.0010)	(0.0010)	(0.0010)	(0.0009)	(0.0009)	(0.0009)	(0.0009)
Year 2010	-0.0017	-0.0017	-0.0017	-0.0017	-0.0018	-0.0018	-0.0018	-0.0018
	(0.0019)	(0.0019)	(0.0019)	(0.0019)	(0.0017)	(0.0017)	(0.0017)	(0.0017)
Year 2011	-0.0013	-0.0013	-0.0013	-0.0013	0.0003	0.0003	0.0003	0.0003
	(0.0021)	(0.0021)	(0.0021)	(0.0021)	(0.0018)	(0.0018)	(0.0018)	(0.0018)
Year 2012	0.0050 *	0.0050 *	0.0050 *	0.0050 *	0.0037	0.0036	0.0037	0.0037
	(0.0024)	(0.0024)	(0.0024)	(0.0024)	(0.0021)	(0.0021)	(0.0021)	(0.0021)
Year 2013	0.0016	0.0016	0.0016	0.0016	-0.0040	-0.0040	-0.0040	-0.0040
	(0.0028)	(0.0028)	(0.0028)	(0.0028)	(0.0024)	(0.0024)	(0.0024)	(0.0024)
Year 2014	0.0060	0.0059	0.0060	0.0059	-0.0001	-0.0001	-0.0001	-0.0001
	(0.0033)	(0.0033)	(0.0033)	(0.0033)	(0.0028)	(0.0028)	(0.0028)	(0.0028)
Year 2015	0.0157 ***	0.0157 ***	0.0157 ***	0.0157 ***	0.0142 ***	0.0142 ***	0.0142 ***	0.0142 ***
	(0.0037)	(0.0037)	(0.0037)	(0.0037)	(0.0031)	(0.0031)	(0.0031)	(0.0031)
Year 2016	0.0233 ***	0.0233 ***	0.0233 ***	0.0233 ***	0.0176 ***	0.0176 ***	0.0176 ***	0.0176 ***
	(0.0041)	(0.0041)	(0.0041)	(0.0041)	(0.0035)	(0.0035)	(0.0035)	(0.0035)
Year 2017	0.0227 ***	0.0227 ***	0.0227 ***	0.0227 ***	0.0162 ***	0.0161 ***	0.0161 ***	0.0161 ***
	(0.0046)	(0.0046)	(0.0046)	(0.0046)	(0.0038)	(0.0038)	(0.0038)	(0.0038)
Year 2018	0.0210 ***	0.0210 ***	0.0210 ***	0.0210 ***	0.0117 **	0.0117 **	0.0117 **	0.0117 **
	(0.0051)	(0.0051)	(0.0051)	(0.0051)	(0.0043)	(0.0043)	(0.0043)	(0.0043)
Year 2019	0.0227 ***	0.0226 ***	0.0227 ***	0.0226 ***	0.0110 *	0.0110 *	0.0109 *	0.0109 *
	(0.0056)	(0.0056)	(0.0056)	(0.0056)	(0.0047)	(0.0047)	(0.0047)	(0.0047)
Ln(N Schools)		0.0136 *		-0.0018		0.0123 *		-0.0013
		(0.0053)		(0.0044)		(0.0047)		(0.0038)
Ln(Avg. Enrollment)		-0.0209 **		-0.0135 *		-0.0221 ***		-0.0142 **
		(0.0068)		(0.0051)		(0.0060)		(0.0044)
% Rural		-0.0340 ***		-0.0115		-0.0396 ***		-0.0180 **
		(0.0066)		(0.0060)		(0.0058)		(0.0052)
White-Black Segregation			0.0179 ***	0.0195 ***			0.0111 *	0.0111 **
			(0.0045)	(0.0046)			(0.0042)	(0.0040)
Free Lunch Segregation			0.0246 ***	0.0211 ***			0.0313 ***	0.0255 ***
			(0.0045)	(0.0051)			(0.0042)	(0.0045)
Within-State SD	0.0212	0.0212	0.0212	0.0212	0.0194	0.0194	0.0194	0.0194
Between-State SD	0.0486	0.0361	0.0282	0.0268	0.0473	0.0321	0.0258	0.0224
Between-State Grade SD	0.0072	0.0072	0.0072	0.0072	0.0062	0.0062	0.0062	0.0062
Between-State Year SD	0.0037	0.0037	0.0037	0.0037	0.0031	0.0031	0.0031	0.0031
R-squared relative to M1/R1	0.0000	0.4492	0.6643	0.6955	0.0000	0.5404	0.7018	0.7750
N Obs.	2942	2942	2942	2942	3079	3079	3079	3079
N States	51	51	51	51	51	51	51	51

Note. ****p* < 0.001; ***p* < 0.01; **p* < 0.05. RLA= Reading/Language Arts. Ln=natural log. Standard errors reported in parentheses. Observations are state-subject-grade-years; groups are states plus DC. R-squared indicates the proportion reduction of between-state intercept variance relative to models M1 and R1. Enrollment, N schools, % Rural, White-Black Segregation, and Free Lunch Segregation are standardized relative to the between-state distribution. Grade and year indicator variables are mean-centered within state and subject.

Table 7

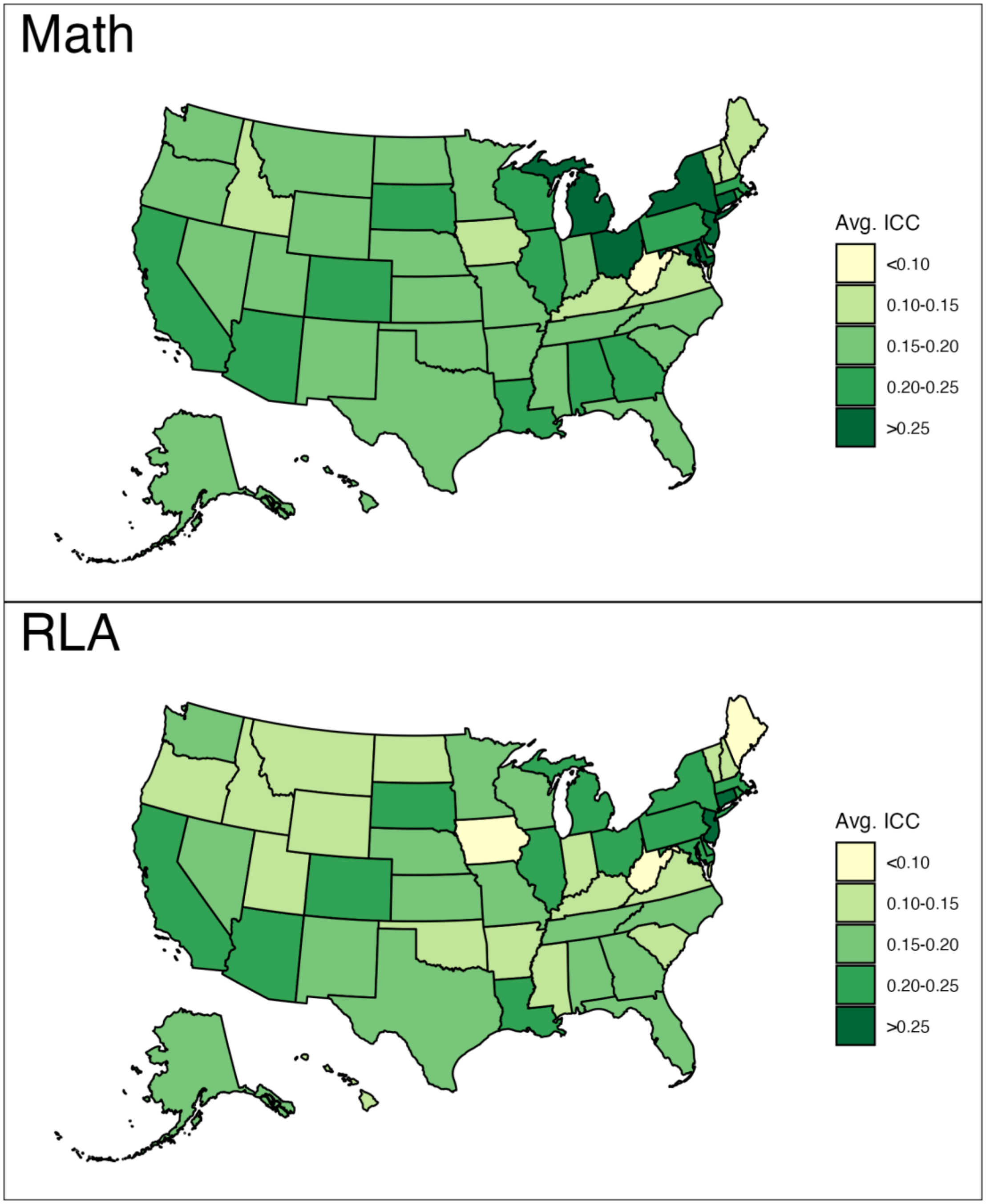
Average ICC Estimates by Subject, Grade, and Year

Subject	**Grade**	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**	**Average**
Math	3	0.180	0.178	0.180	0.186	0.187	0.191	0.197	0.199	0.195	0.198	0.197	*0.190*
	4	0.180	0.178	0.182	0.190	0.190	0.194	0.209	0.219	0.211	0.213	0.211	*0.198*
	5	0.188	0.189	0.186	0.192	0.195	0.200	0.214	0.217	0.219	0.213	0.215	*0.202*
	6	0.185	0.184	0.184	0.190	0.187	0.187	0.198	0.208	0.208	0.210	0.210	*0.196*
	7	0.185	0.186	0.183	0.193	0.183	0.178	0.174	0.180	0.188	0.186	0.188	*0.184*
	8	0.194	0.193	0.192	0.203	0.189	0.191	0.192	0.195	0.193	0.191	0.198	*0.194*
	Average	*0.185*	*0.184*	*0.184*	*0.192*	*0.189*	*0.190*	*0.197*	*0.203*	*0.202*	*0.202*	*0.203*	*0.194*
RLA	3	0.157	0.158	0.157	0.168	0.164	0.167	0.185	0.186	0.183	0.183	0.185	*0.172*
	4	0.163	0.161	0.163	0.171	0.165	0.168	0.184	0.194	0.184	0.182	0.186	*0.175*
	5	0.166	0.163	0.163	0.170	0.165	0.171	0.181	0.184	0.187	0.180	0.184	*0.174*
	6	0.163	0.161	0.160	0.165	0.155	0.153	0.167	0.175	0.175	0.171	0.170	*0.165*
	7	0.162	0.163	0.163	0.165	0.152	0.153	0.163	0.168	0.169	0.160	0.163	*0.162*
	8	0.162	0.162	0.167	0.168	0.154	0.150	0.164	0.159	0.160	0.157	0.161	*0.160*
	Average	*0.162*	*0.161*	*0.162*	*0.168*	*0.159*	*0.160*	*0.174*	*0.178*	*0.176*	*0.172*	*0.175*	*0.168*

Note. ICC=intraclass correlation coefficient. RLA= Reading/Language Arts. Row and column averages, presented in italics, are unweighted across cells.

Figure 1

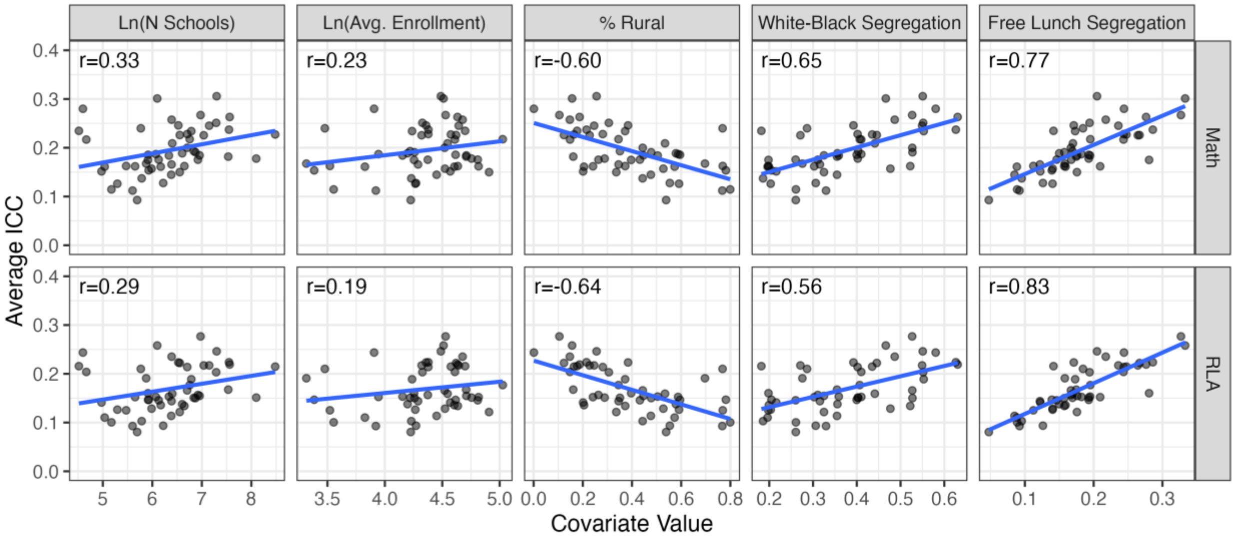
Map of Average ICC Estimates Across States



Note. ICC=intraclass correlation coefficient. RLA= Reading/Language Arts. Average ICC estimates are Empirical Bayes (EB) estimates from models M1 and R1 in Table 6.

Figure 2

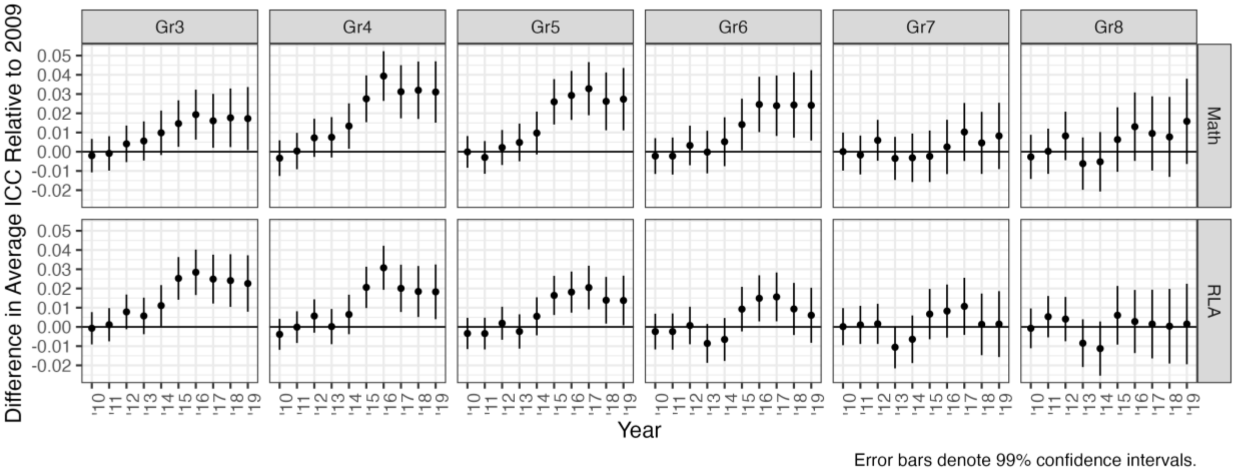
Average ICC Estimates Versus State Covariates



Note. ICC=intraclass correlation coefficient. RLA= Reading/Language Arts. Average ICC estimates are Empirical Bayes (EB) estimates from models M1 and R1 in Table 6.

Figure 3

Changes in Average ICC Estimates Across Years, by Subject and Grade



Note. ICC=intraclass correlation coefficient. RLA= Reading/Language Arts. Point estimates are derived from models fit separately to each grade and subject. Models include state random intercepts and state random linear year trends, as described in text.

Supplementary Materials for:

Shear, B. R., Taylor, J. A., & Fahle, E. M. (2026). Determinants of between-school variation in student achievement: Results from US population data. *Journal of Research on Educational Effectiveness*, 1–25. https://doi.org/10.1080/19345747.2025.2598314

Table S1

State-Average Math ICCs by Grade

	Grade					
State	**3**	**4**	**5**	**6**	**7**	**8**
AK	0.187	0.183	0.154	0.175	0.160	0.147
AL	0.191	0.195	0.213	0.216	0.212	0.227
AR	0.173	0.166	0.179	0.142	0.134	0.144
AZ	0.213	0.216	0.224	0.233	0.233	0.239
CA	0.207	0.222	0.243	0.219	0.242	0.250
CO	0.224	0.233	0.234	0.223	0.223	0.225
CT	0.274	0.296	0.302	0.309	0.304	0.324
DC	0.288	0.290	0.278	0.240	0.273	0.324
DE	0.222	0.229	0.223	0.226	0.227	0.281
FL	0.170	0.165	0.172	0.223	0.171	0.186
GA	0.203	0.219	0.230	0.208	0.213	0.232
HI	0.167	0.155	0.155	0.165	0.127	0.139
IA	0.141	0.149	0.156	0.126	0.096	0.097
ID	0.124	0.136	0.134	0.152	0.136	0.141
IL	0.252	0.252	0.252	0.236	0.216	0.216
IN	0.168	0.165	0.172	0.161	0.148	0.159
KS	0.194	0.197	0.197	0.187	0.162	0.167
KY	0.146	0.153	0.154	0.139	0.129	0.146
LA	0.233	0.251	0.247	0.236	0.241	0.270
MA	0.211	0.214	0.234	0.223	0.239	0.236
MD	0.228	0.242	0.256	0.247	0.297	0.313
ME	0.119	0.113	0.119	0.110	0.102	0.109
MI	0.230	0.260	0.274	0.251	0.243	0.249
MN	0.197	0.205	0.199	0.206	0.178	0.173
MO	0.198	0.212	0.218	0.186	0.152	0.157
MS	0.196	0.198	0.200	0.177	0.161	0.181
MT	0.171	0.166	0.169	0.144	0.137	0.134
NC	0.168	0.177	0.179	0.187	0.189	0.205
ND	0.149	0.156	0.151	0.183	0.176	0.160
NE	0.198	0.224	0.213	0.184	0.150	0.154
NH	0.135	0.139	0.135	0.118	0.109	0.120
NJ	0.253	0.259	0.275	0.269	0.276	0.278
NM	0.172	0.185	0.172	0.191	0.159	0.167
NV	0.135	0.148	0.156	0.180	0.167	0.184
NY	0.229	0.246	0.256	0.278	0.297	0.286

OH	0.272	0.314	0.329	0.335	0.284	0.281
OK	0.208	0.207	0.205	0.184	0.153	0.172
OR	0.175	0.177	0.182	0.153	0.153	0.153
PA	0.235	0.253	0.270	0.251	0.231	0.234
RI	0.203	0.196	0.194	0.195	0.251	0.268
SC	0.168	0.187	0.185	0.171	0.172	0.172
SD	0.254	0.258	0.237	0.230	0.230	0.231
TN	0.179	0.202	0.216	0.201	0.201	NA
TX	0.162	0.177	0.180	0.200	0.158	0.167
UT	0.143	0.148	0.164	0.162	0.187	0.176
VA	0.141	0.145	0.158	0.205	NA	NA
VT	0.108	0.119	0.127	0.106	0.103	0.122
WA	0.172	0.192	0.182	0.181	0.158	0.160
WI	0.221	0.236	0.249	0.239	0.228	0.233
WV	0.101	0.098	0.108	0.086	0.073	0.088
WY	0.163	0.162	0.173	0.172	0.137	0.160

Note. ICC=intraclass correlation coefficient. The average ICCs shown in the table are unweighted; averages are within state and grade across years where estimates are available. Sample and estimation details provided in main text.

Table S2

State-Average RLA ICCs by Grade

	Grade					
State	**3**	**4**	**5**	**6**	**7**	**8**
AK	0.187	0.183	0.154	0.175	0.160	0.147
AL	0.191	0.195	0.213	0.216	0.212	0.227
AR	0.173	0.166	0.179	0.142	0.134	0.144
AZ	0.213	0.216	0.224	0.233	0.233	0.239
CA	0.207	0.222	0.243	0.219	0.242	0.250
CO	0.224	0.233	0.234	0.223	0.223	0.225
CT	0.274	0.296	0.302	0.309	0.304	0.324
DC	0.288	0.290	0.278	0.240	0.273	0.324
DE	0.222	0.229	0.223	0.226	0.227	0.281
FL	0.170	0.165	0.172	0.223	0.171	0.186
GA	0.203	0.219	0.230	0.208	0.213	0.232
HI	0.167	0.155	0.155	0.165	0.127	0.139
IA	0.141	0.149	0.156	0.126	0.096	0.097
ID	0.124	0.136	0.134	0.152	0.136	0.141
IL	0.252	0.252	0.252	0.236	0.216	0.216
IN	0.168	0.165	0.172	0.161	0.148	0.159
KS	0.194	0.197	0.197	0.187	0.162	0.167
KY	0.146	0.153	0.154	0.139	0.129	0.146
LA	0.233	0.251	0.247	0.236	0.241	0.270
MA	0.211	0.214	0.234	0.223	0.239	0.236
MD	0.228	0.242	0.256	0.247	0.297	0.313
ME	0.119	0.113	0.119	0.110	0.102	0.109
MI	0.230	0.260	0.274	0.251	0.243	0.249
MN	0.197	0.205	0.199	0.206	0.178	0.173
MO	0.198	0.212	0.218	0.186	0.152	0.157
MS	0.196	0.198	0.200	0.177	0.161	0.181
MT	0.171	0.166	0.169	0.144	0.137	0.134
NC	0.168	0.177	0.179	0.187	0.189	0.205
ND	0.149	0.156	0.151	0.183	0.176	0.160
NE	0.198	0.224	0.213	0.184	0.150	0.154
NH	0.135	0.139	0.135	0.118	0.109	0.120
NJ	0.253	0.259	0.275	0.269	0.276	0.278
NM	0.172	0.185	0.172	0.191	0.159	0.167
NV	0.135	0.148	0.156	0.180	0.167	0.184
NY	0.229	0.246	0.256	0.278	0.297	0.286

OH	0.272	0.314	0.329	0.335	0.284	0.281
OK	0.208	0.207	0.205	0.184	0.153	0.172
OR	0.175	0.177	0.182	0.153	0.153	0.153
PA	0.235	0.253	0.270	0.251	0.231	0.234
RI	0.203	0.196	0.194	0.195	0.251	0.268
SC	0.168	0.187	0.185	0.171	0.172	0.172
SD	0.254	0.258	0.237	0.230	0.230	0.231
TN	0.179	0.202	0.216	0.201	0.201	NA
TX	0.162	0.177	0.180	0.200	0.158	0.167
UT	0.143	0.148	0.164	0.162	0.187	0.176
VA	0.141	0.145	0.158	0.205	NA	NA
VT	0.108	0.119	0.127	0.106	0.103	0.122
WA	0.172	0.192	0.182	0.181	0.158	0.160
WI	0.221	0.236	0.249	0.239	0.228	0.233
WV	0.101	0.098	0.108	0.086	0.073	0.088
WY	0.163	0.162	0.173	0.172	0.137	0.160

Note. ICC=intraclass correlation coefficient. RLA=reading/language arts. The average ICCs shown in the table are unweighted; averages are within state and grade across years where estimates are available. Sample and estimation details provided in main text.

Table S3

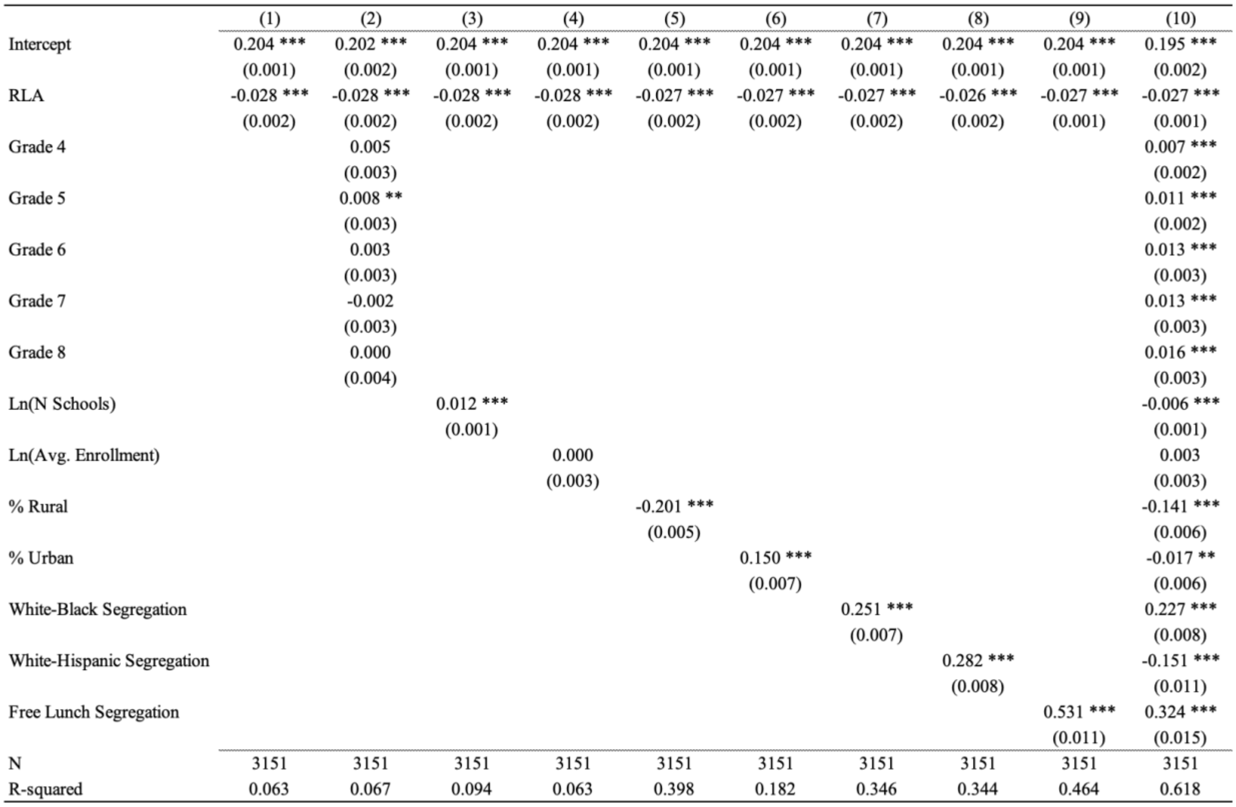
Descriptive Statistics and Correlations Among School Characteristics and ICC Estimates, 80% Sample

Row	Variable	N	M	SD	Min	Max	1	2	3	4	5	6	7	8	9	10	11
1	ICC Math	42	0.20	0.05	0.09	0.31	1.00										
2	ICC RLA	41	0.18	0.05	0.09	0.28	0.93	1.00									
3	N Schools	42	930.71	905.06	95.26	5021.15	0.21	0.25	1.00								
4	Avg. Enrollment	42	95.47	23.58	50.83	162.92	-0.05	-0.07	0.14	1.00							
5	ln(N Schools)	42	6.45	0.87	4.52	8.48	0.26	0.20	0.83	0.11	1.00						
6	ln(Enrollment)	42	4.49	0.23	3.91	5.03	-0.02	-0.04	0.20	0.98	0.19	1.00					
7	% Rural	42	0.33	0.16	0.00	0.60	-0.56	-0.68	-0.26	-0.22	-0.07	-0.21	1.00				
8	% Urban	42	0.27	0.14	0.08	1.00	0.44	0.42	0.19	-0.15	-0.01	-0.21	-0.58	1.00			
9	H White-Black	42	0.40	0.13	0.18	0.63	0.67	0.57	0.34	-0.06	0.45	-0.03	-0.27	0.38	1.00		
10	H White-Hispanic	42	0.30	0.10	0.14	0.56	0.62	0.68	0.33	0.02	0.25	0.00	-0.61	0.58	0.71	1.00	
11	H Free Lunch	42	0.19	0.06	0.05	0.33	0.71	0.78	0.29	0.08	0.28	0.10	-0.62	0.41	0.48	0.72	1.00

Note. ICC=intraclass correlation coefficient. RLA=reading/language arts. Ln=natural log. Observations are states; data are averages across grades and years within states. Sample is restricted to the set of states for which sample coverage for number of schools was at least 80%. Observations are states plus DC with available data. Variables are averages across grades and years within states. School structure covariates are calculated for all grades and years, not only grades and years with ICC estimates.

Table S4

OLS Regression Model Estimates, 80% Sample



	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Intercept	0.204 ***	0.202 ***	0.204 ***	0.204 ***	0.204 ***	0.204 ***	0.204 ***	0.204 ***	0.204 ***	0.195 ***
	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)
RLA	-0.028 ***	-0.028 ***	-0.028 ***	-0.028 ***	-0.027 ***	-0.027 ***	-0.027 ***	-0.026 ***	-0.027 ***	-0.027 ***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)
Grade 4		0.005								0.007 ***
		(0.003)								(0.002)
Grade 5		0.008 **								0.011 ***
		(0.003)								(0.002)
Grade 6		0.003								0.013 ***
		(0.003)								(0.003)
Grade 7		-0.002								0.013 ***
		(0.003)								(0.003)
Grade 8		0.000								0.016 ***
		(0.004)								(0.003)
Ln(N Schools)			0.012 ***							-0.006 ***
			(0.001)							(0.001)
Ln(Avg. Enrollment)				0.000						0.003
				(0.003)						(0.003)
% Rural					-0.201 ***					-0.141 ***
					(0.005)					(0.006)
% Urban						0.150 ***				-0.017 **
						(0.007)				(0.006)
White-Black Segregation							0.251 ***			0.227 ***
							(0.007)			(0.008)
White-Hispanic Segregation								0.282 ***		-0.151 ***
								(0.008)		(0.011)
Free Lunch Segregation									0.531 ***	0.324 ***
									(0.011)	(0.015)
N	3151	3151	3151	3151	3151	3151	3151	3151	3151	3151
R-squared	0.063	0.067	0.094	0.063	0.398	0.182	0.346	0.344	0.464	0.618

Note. ****p* < 0.001; ***p* < 0.01; **p* < 0.05. RLA=reading/language arts. Ln=natural log. Observations are state-grade-subject-years. All covariates are grand mean centered except subject and grade fixed effects. Sample is restricted to the set of ICC estimates for which sample coverage was at least 80% of schools.

Table S5

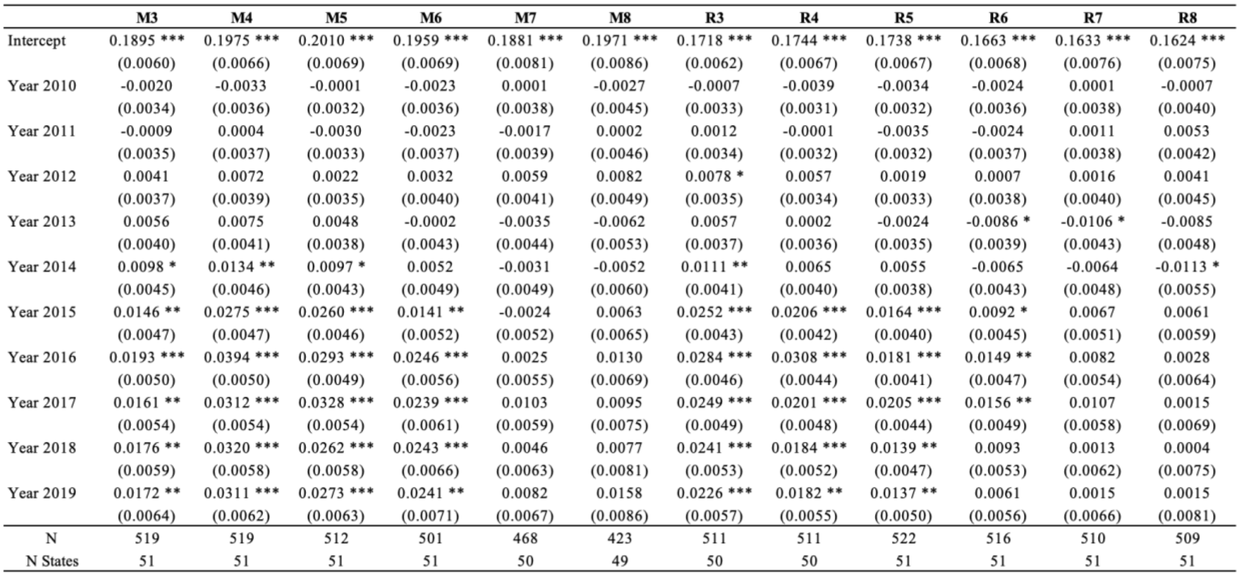
HLM Regression Model Estimates, 80% Sample

	Math				RLA			
	M1	M2	M3	M4	R1	R2	R3	R4
Intercept	0.2001 ***	0.1957 ***	0.1932 ***	0.1947 ***	0.1773 ***	0.1719 ***	0.1695 ***	0.1690 ***
	(0.0075)	(0.0071)	(0.0046)	(0.0052)	(0.0069)	(0.0058)	(0.0041)	(0.0043)
Grade	0.0032 **	0.0032 **	0.0034 **	0.0033 **	-0.0019	-0.0021	-0.0019	-0.0019
	(0.0012)	(0.0012)	(0.0012)	(0.0012)	(0.0011)	(0.0011)	(0.0011)	(0.0011)
Year 2010	-0.0013	-0.0013	-0.0012	-0.0012	-0.0009	-0.0009	-0.0009	-0.0009
	(0.0020)	(0.0020)	(0.0020)	(0.0020)	(0.0018)	(0.0018)	(0.0018)	(0.0018)
Year 2011	0.0013	0.0013	0.0014	0.0013	0.0025	0.0025	0.0025	0.0025
	(0.0023)	(0.0023)	(0.0023)	(0.0023)	(0.0020)	(0.0020)	(0.0020)	(0.0020)
Year 2012	0.0084 **	0.0083 **	0.0085 **	0.0085 **	0.0071 **	0.0072 **	0.0072 **	0.0072 **
	(0.0026)	(0.0026)	(0.0026)	(0.0026)	(0.0024)	(0.0024)	(0.0024)	(0.0024)
Year 2013	0.0081 **	0.0080 **	0.0082 **	0.0082 **	0.0014	0.0014	0.0014	0.0014
	(0.0030)	(0.0030)	(0.0030)	(0.0030)	(0.0028)	(0.0028)	(0.0028)	(0.0028)
Year 2014	0.0131 ***	0.0129 ***	0.0132 ***	0.0131 ***	0.0062	0.0062	0.0063	0.0062
	(0.0036)	(0.0036)	(0.0036)	(0.0036)	(0.0034)	(0.0034)	(0.0034)	(0.0034)
Year 2015	0.0230 ***	0.0227 ***	0.0231 ***	0.0229 ***	0.0200 ***	0.0200 ***	0.0200 ***	0.0200 ***
	(0.0041)	(0.0041)	(0.0041)	(0.0041)	(0.0039)	(0.0039)	(0.0039)	(0.0039)
Year 2016	0.0276 ***	0.0273 ***	0.0277 ***	0.0275 ***	0.0198 ***	0.0198 ***	0.0198 ***	0.0198 ***
	(0.0046)	(0.0046)	(0.0046)	(0.0046)	(0.0044)	(0.0044)	(0.0044)	(0.0044)
Year 2017	0.0275 ***	0.0271 ***	0.0276 ***	0.0274 ***	0.0190 ***	0.0190 ***	0.0190 ***	0.0190 ***
	(0.0051)	(0.0051)	(0.0051)	(0.0051)	(0.0049)	(0.0049)	(0.0049)	(0.0049)
Year 2018	0.0272 ***	0.0268 ***	0.0273 ***	0.0271 ***	0.0160 **	0.0160 **	0.0161 **	0.0160 **
	(0.0056)	(0.0056)	(0.0056)	(0.0056)	(0.0054)	(0.0054)	(0.0054)	(0.0054)
Year 2019	0.0278 ***	0.0273 ***	0.0279 ***	0.0277 ***	0.0180 **	0.0180 **	0.0181 **	0.0180 **
	(0.0061)	(0.0061)	(0.0061)	(0.0061)	(0.0059)	(0.0059)	(0.0059)	(0.0059)
Ln(N Schools)		0.0085		-0.0058		0.0092		-0.0015
		(0.0058)		(0.0047)		(0.0046)		(0.0039)
Ln(Avg. Enrollment)		-0.0208 *		-0.0109		-0.0196 *		-0.0099
		(0.0103)		(0.0075)		(0.0081)		(0.0062)
% Rural		-0.0357 ***		-0.0124		-0.0359 ***		-0.0191 **
		(0.0077)		(0.0071)		(0.0060)		(0.0058)
White-Black Segregation			0.0204 ***	0.0215 ***			0.0129 **	0.0132 **
			(0.0051)	(0.0052)			(0.0045)	(0.0043)
Free Lunch Segregation			0.0241 ***	0.0196 **			0.0273 ***	0.0191 ***
			(0.0051)	(0.0059)			(0.0045)	(0.0048)
Within-State SD	0.0177	0.0177	0.0177	0.0177	0.0154	0.0154	0.0154	0.0154
Between-State SD	0.0487	0.0382	0.0289	0.0277	0.0443	0.0306	0.0254	0.0221
Between-State Grade SD	0.0068	0.0069	0.0069	0.0069	0.0064	0.0064	0.0065	0.0065
Between-State Year SD	0.0035	0.0035	0.0035	0.0035	0.0034	0.0034	0.0034	0.0034
R-squared relative to M1	0.0000	0.3828	0.6480	0.6762	0.0000	0.5221	0.6705	0.7514
N Obs.	1587	1587	1587	1587	1564	1564	1564	1564
N States	42	42	42	42	41	41	41	41

Note. ***p < 0.001; **p < 0.01; *p < 0.05. RLA=reading/language arts. Ln=natural log. Standard errors reported in parentheses. Observations are state-subject-grade-years; groups are states plus DC. R-squared indicates the proportion reduction of between-state intercept variance relative to model 1. Enrollment, N schools, % Rural, White-Black Segregation, and Free Lunch Segregation are standardized relative to the between-state distribution. Grade and year indicator variables are mean-centered within state and subject. Sample is restricted to the set of ICC estimates for which sample coverage was at least 80% of schools.

Table S6

HLM Year-by-Grade Regression Models, Full Sample



	M3	M4	M5	M6	M7	M8	R3	R4	R5	R6	R7	R8
Intercept	0.1895 ***	0.1975 ***	0.2010 ***	0.1959 ***	0.1881 ***	0.1971 ***	0.1718 ***	0.1744 ***	0.1738 ***	0.1663 ***	0.1633 ***	0.1624 ***
	(0.0060)	(0.0066)	(0.0069)	(0.0069)	(0.0081)	(0.0086)	(0.0062)	(0.0067)	(0.0067)	(0.0068)	(0.0076)	(0.0075)
Year 2010	-0.0020	-0.0033	-0.0001	-0.0023	0.0001	-0.0027	-0.0007	-0.0039	-0.0034	-0.0024	0.0001	-0.0007
	(0.0034)	(0.0036)	(0.0032)	(0.0036)	(0.0038)	(0.0045)	(0.0033)	(0.0031)	(0.0032)	(0.0036)	(0.0038)	(0.0040)
Year 2011	-0.0009	0.0004	-0.0030	-0.0023	-0.0017	0.0002	0.0012	-0.0001	-0.0035	-0.0024	0.0011	0.0053
	(0.0035)	(0.0037)	(0.0033)	(0.0037)	(0.0039)	(0.0046)	(0.0034)	(0.0032)	(0.0032)	(0.0037)	(0.0038)	(0.0042)
Year 2012	0.0041	0.0072	0.0022	0.0032	0.0059	0.0082	0.0078 *	0.0057	0.0019	0.0007	0.0016	0.0041
	(0.0037)	(0.0039)	(0.0035)	(0.0040)	(0.0041)	(0.0049)	(0.0035)	(0.0034)	(0.0033)	(0.0038)	(0.0040)	(0.0045)
Year 2013	0.0056	0.0075	0.0048	-0.0002	-0.0035	-0.0062	0.0057	0.0002	-0.0024	-0.0086 *	-0.0106 *	-0.0085
	(0.0040)	(0.0041)	(0.0038)	(0.0043)	(0.0044)	(0.0053)	(0.0037)	(0.0036)	(0.0035)	(0.0039)	(0.0043)	(0.0048)
Year 2014	0.0098 *	0.0134 **	0.0097 *	0.0052	-0.0031	-0.0052	0.0111 **	0.0065	0.0055	-0.0065	-0.0064	-0.0113 *
	(0.0045)	(0.0046)	(0.0043)	(0.0049)	(0.0049)	(0.0060)	(0.0041)	(0.0040)	(0.0038)	(0.0043)	(0.0048)	(0.0055)
Year 2015	0.0146 **	0.0275 ***	0.0260 ***	0.0141 **	-0.0024	0.0063	0.0252 ***	0.0206 ***	0.0164 ***	0.0092 *	0.0067	0.0061
	(0.0047)	(0.0047)	(0.0046)	(0.0052)	(0.0052)	(0.0065)	(0.0043)	(0.0042)	(0.0040)	(0.0045)	(0.0051)	(0.0059)
Year 2016	0.0193 ***	0.0394 ***	0.0293 ***	0.0246 ***	0.0025	0.0130	0.0284 ***	0.0308 ***	0.0181 ***	0.0149 **	0.0082	0.0028
	(0.0050)	(0.0050)	(0.0049)	(0.0056)	(0.0055)	(0.0069)	(0.0046)	(0.0044)	(0.0041)	(0.0047)	(0.0054)	(0.0064)
Year 2017	0.0161 **	0.0312 ***	0.0328 ***	0.0239 ***	0.0103	0.0095	0.0249 ***	0.0201 ***	0.0205 ***	0.0156 **	0.0107	0.0015
	(0.0054)	(0.0054)	(0.0054)	(0.0061)	(0.0059)	(0.0075)	(0.0049)	(0.0048)	(0.0044)	(0.0049)	(0.0058)	(0.0069)
Year 2018	0.0176 **	0.0320 ***	0.0262 ***	0.0243 ***	0.0046	0.0077	0.0241 ***	0.0184 ***	0.0139 **	0.0093	0.0013	0.0004
	(0.0059)	(0.0058)	(0.0058)	(0.0066)	(0.0063)	(0.0081)	(0.0053)	(0.0052)	(0.0047)	(0.0053)	(0.0062)	(0.0075)
Year 2019	0.0172 **	0.0311 ***	0.0273 ***	0.0241 **	0.0082	0.0158	0.0226 ***	0.0182 **	0.0137 **	0.0061	0.0015	0.0015
	(0.0064)	(0.0062)	(0.0063)	(0.0071)	(0.0067)	(0.0086)	(0.0057)	(0.0055)	(0.0050)	(0.0056)	(0.0066)	(0.0081)
N	519	519	512	501	468	423	511	511	522	516	510	509
N States	51	51	51	51	50	49	50	50	51	51	51	51

Note. *** p < 0.001; ** p < 0.01; * p < 0.05. Observations are state-subject-grade-years; groups are states. Standard errors reported in parentheses. Each column (model) is for a single grade-by-subject combination (M3=grade 3 math, M4=grade 4 math, R3=grade 3 RLA, etc.). Models include random state intercepts and randomly varying linear year trends across states. Each coefficient represents the average within-state difference in intraclass correlation coefficient (ICC) relative to 2009 (the omitted year).