Title: Between-District Test Score Variation, 2009-2012

Authors and Affiliations:

Erin Fahle, Stanford University
Sean Reardon, Stanford University

Background / Context:

Describing the variation in test scores between and within school districts is critical for (1) for policy-related and descriptive work that investigates the sorting of students among districts and the differential effectiveness of those districts; and (2) for methodological work planning future experiments or interventions (Hedges & Hedberg 2007, 2014; Jacob et al. 2010; Westine et al. 2013). Intraclass Correlations (ICCs) and Coefficients of Variation (CVs) are two complementary ways to describe test score variation. ICCs describe the proportion of variance in test scores that is between (rather than within) school districts or schools. CVs describe the extent of heteroscedasticity in district (or school) test score distributions. The most straightforward method of calculating ICCs and CVs is to use student-level data to directly estimate the means and variances of district or school test score distributions.

Most prior research on ICCs aligns with the second motivation and uses 2-Level (students in schools/classrooms nested in districts/states) or 3-Level (students in schools/classrooms nested in districts in states) Hierarchical Linear Modeling (HLM) to estimate ICCs using data from U.S. schools, districts, or states or data from nationally representative studies. Contributors to this literature include Hedges and Hedberg (2007, 2014), Xu and Nichols (2010), Bloom, Richburg-Hayes, and Black (2007), Jacob, Zhu, and Bloom (2010), Scochet (2008), Westine, Spybrook and Taylor (2013), Zhu, Jacob, Bloom and Xu (2012), Konstantopoulos (2009, 2011, 2012), and Brandon, Harrison and Lawton (2013). Little published research focused on estimating CVs is available.

However, student-level data for the full population of schools or districts are typically not readily available to researchers. Therefore, ICC and CV estimates are not readily available in the literature for many populations of interest (e.g., most U.S. states) for either methodological or policy-related work. The most comprehensive paper to date is Hedges and Hedberg (2014), which uses data from 11 states to compute between-district and -school ICCs for one year (2009-10 in most cases). Hedges and Hedberg find that there is substantial variation in ICCs across states and subjects. Comparing to their prior work, the authors show that the state-level results are not well-summarized by the national ICCs calculated in Hedges and Hedberg (2007), motivating the need to investigate state-level ICCs systematically.

Purpose / Objective / Research Question / Focus of Study:

In this study we investigate three interrelated research questions. First, how much between-district variation exists across U.S. states? Second, what are they key patterns across grades, subject, and years (within or across states) in the between-district variation? And, third what state-level factors are associated with a state having more-or-less observed between-district variation? To address these questions, we leverage recent developments in ordered probit models to estimate two measure of heteroscedasticity, the between-district ICCs and the state-level CVs, for 49 states across four years, six grades and two subjects from coarsened test score data. The availability of estimates across nearly all states for multiple grades and years enables a systematic analysis of the patterns in the ICCs and CVs, and, further enables modeling the relationships between those estimated measures of heteroscedasticity and state-level characteristics to better understand what may shape between-district differences.

Significance / Novelty of study:

Due to the constraints of standard methods for calculating ICCs the CVs (i.e. the need for student-level data), prior work has been unable to provide a comprehensive set of estimated ICCs and CVs for further methodological or policy-related analyses. This work leverages recent developments in ordered probit models, which enable us to use ordinal proficiency data (which is more readily available than student-level data) to estimate between-district ICCs and CVs in 49 states, in grades 3 through 8, in both math and reading, in four school years (2008-09 through 2011-12). This work not only substantially increases the number of available ICC and CV estimates in the literature, but also begins the policy-related, descriptive analysis through modeling patterns in the ICCs and CVs, as well as their relationship between state-level factors that may influence between-district differences across the U.S. To our knowledge, no prior paper has systematically analyzed ICCs and CVs across the U.S. (particularly across all states, in multiple years) in this way.

Statistical, Measurement, or Econometric Model:

This study leverages the use of ordered probit models to recover distributional information from coarsened test score data proposed by Reardon et al. (2015). The authors focus the paper on the heteroscedastic ordered probit (HETOP) model, demonstrating the ability to recover group means and standard deviations (SDs), and using these parameters to subsequently estimate ICCs and CVs. The simulation and real data application results demonstrate that the HETOP model provides ICC estimates that are very slightly positively biased (the underestimation of within-district variance leads to a corresponding overestimation of the proportion of variance that is between districts). This bias is evident only when group sample sizes are small and when the thresholds defining the ordered categories are highly skewed or widely spaced; in most cases that the authors studied, the ICC bias was generally less than 0.01. Because district-level test score data often yield small group sizes, we adopt their recommendation to use a partially heteroscedastic ordered probit (PHOP) model as it reduces the RMSE relative to the HETOP model.

Adopting the notation from Reardon et al. (2015), the calculation of the ICC, sampling variance of the ICC, and CV are as follows:[*]

Let:

$\mathbf{P} = [p_1, p_2, \dots p_K]$ be the $1 \times K$ vector of districts' population proportions, so that $\sum_k p_k = 1$;
$\widehat{\mathbf{\Sigma}}^* = [\hat{\sigma}_1^*, \hat{\sigma}_2^*, \dots \hat{\sigma}_K^*]^t$ be the $K \times 1$ vector of districts' estimated standard deviations;
$\widehat{\mathbf{M}}^* = [\hat{\mu}_1^*, \hat{\mu}_2^* \dots \hat{\mu}_K^*]^t$ be the K×1 vector of districts' estimated mean.

We then can write the ICC as:

$$\widehat{ICC} = 1 - \mathbf{P}\left(\widehat{\mathbf{\Sigma}}^*\right)^2,$$

and the sampling variance of the ICC as:

$$Var\left(\widehat{ICC}\right) = Var\left(\mathbf{P}\left(\widehat{\mathbf{\Sigma}}^*\right)^2\right) = 4\mathbf{P}\widehat{\mathbf{\Sigma}}^{*d}\widehat{\mathbf{W}}^*\widehat{\mathbf{\Sigma}}^{*d},$$

where $\widehat{\mathbf{W}}^*$ is the estimated variance-covariance matrix of the SDs.

[*] Note $[\quad]^t$ indicates the transpose of a matrix; $[\quad]^2$ indicates the Hadamard product of a matrix; and $[\quad]^d$ indicates a $K \times K$ diagonal matrix whose diagonal entry x_k is the expected value of $\hat{x}_k$.

It is important to note is that this approach provides an estimate of the between-district proportion of the variance in test scores. This estimate will differ from that recovered using alternative methods such as HLM which estimate the variance in the district means. In the paper we discuss the differences in these estimands and their potential uses.

To calculate the CV of the SDs and variances, we use the estimated district SDs $\left(\hat{\sigma}_g^*\right)$ and their sampling variances $\left(\hat{v}_g^*\right)$ to fit the precision-weighted random-effects model:

$$\hat{\sigma}_g^* = \beta_g + e_g$$
$$\beta_g = \gamma + u_g$$
$$u_g \sim N(0, \tau^2); e_g \sim N\left(0, \hat{v}_g^*\right).$$

From this model, we obtain estimates of γ, the average of the true group SDs, and τ^2, the variance of the true group SDs; we then estimate the CV of the SDs as:

$$\widehat{CV}(\sigma) = \frac{\hat{\tau}}{\hat{\gamma}}$$

Once we have estimated $CV(\sigma)$, we can also compute an estimate of CV of the variances ($CV(\sigma^2)$) by noting that:

$$CV(\sigma^2) = \frac{\sqrt{var(\sigma^2)}}{E(\sigma^2)} \approx \frac{\sqrt{4E[\sigma]^2 var(\sigma)}}{E[\sigma]^2 + var(\sigma)} = \frac{\sqrt{4\gamma^2\tau^2}}{\gamma^2 + \tau^2} = \frac{2\gamma\tau}{\gamma^2 + \tau^2} = \frac{2CV(\sigma)}{1 + CV(\sigma)^2}.$$

Usefulness / Applicability of Method:

We apply this method to a large data set provided by the National Center of Education Statistics (NCES) through a restricted data license. Under the No Child Left Behind (NCLB) legislation, states are required to report aggregated test score results to the U.S. Department of Education, through a system called EdFacts. These data include counts of students in each of several ordered proficiency categories (labeled, for example, as “below basic,” “basic,” “proficient,” and “advanced”), by school district, year, grade, and test subject. We use a subset of the data for 49 states across four years (2009 – 2012) for six grades (3rd – 8th) in two subjects (ELA and mathematics). Note that we exclude Hawai’i and the District of Columbia because both have only a single school district. For our analysis, we define a school district as a Local Education Agency (LEA) that serves students in at least one of grades 3-8, which includes both elementary and unified districts. We include charter schools as part of the public school district in which they are chartered, or, if they are not affiliated with a traditional public school district, we assign them to the public district in which they are geographically located.

Findings / Results:

We find substantial variation in ICCs and CVs across states. The state average between district ICC estimates range from <0.01 to 0.22, and the state average CVs of the variances range from 0.10 to 0.25. Within states, ICCs are very highly correlated across subjects and nearly identical across years (please insert table 1 here). The CVs vary slightly more within states across grades and years, and are less correlated across subjects. Both ICCs and CVs are generally larger in math than in reading, and larger in later grades than earlier grades (particularly in math). Using the Common Core of Data (CCD), we further investigate the association between ICCs and features of states’ educational systems to explain the differences in ICCs observed across states (please insert figure 1 here). For each state, we calculate the average number of districts (that serve a grade-level, across years), the average district enrollment (within a grade, across years), and the standard deviation of districts' grade-level enrollments. Additionally, we calculate the Herfindahl Index as a measure of the concentration of students across districts.

We find that the number of districts, the mean enrollment across districts, and the standard deviation of enrollment across districts within a state are not highly correlated with the ICCs (results not shown). However, both economic and racial segregation are strongly correlated with ICCs (please insert figure 2 here). There are two potential explanations for this finding. First, it may reflect the influence of student's family socioeconomic status (income, parental education, etc.) on academic performance. Because socioeconomic status affects students' academic performance, states with higher levels of between-district economic segregation will likely have more between-district variation in test scores (i.e. larger ICCs). A second possibility is that, high levels of segregation may lead to large between-district differences in school quality, which may lead, in turn, to large differences in academic performance. In this case, differences in school quality—rather than differences in family background—are the mechanism through which segregation is related to the ICC. The two explanations are not mutually exclusive, of course. From our data, we cannot determine the contribution of each mechanism to the pattern we observe. Nonetheless, this finding warrants future research.

Conclusions:

The ordinal probit model enables us to estimate between-district ICCs and CVs from ordinal proficiency data, which are more readily available than student-level data. In this work, we use these methods to estimate ICCs and CVs in each state across multiple grades, years and subject, substantially increasing the number of ICC and CV estimates in the literature. We study the patterns in the ICC and CV estimates across grades, years and subject within states, finding clear patterns, such as the evidence that ICCs are larger in math than reading, that merit future investigation as to factors that may drive these patterns. We further investigate and find critical factors related to the average between-district variation in test scores across states. This work serves as a first step in understanding what shapes between-district variation in achievement. Moreover, the state- and grade-specific estimates of ICCs may be useful for designing sampling and or randomization plan for surveys and evaluation studies.

Appendices

Appendix A. References

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. doi: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. doi:10.1177/1098214012466453

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. doi:10.3102/0162373707299706

Hedges, L. V., & Hedberg, E. C. (2013). Intraclass Correlations and Covariate Outcome Correlations for Planning Two- and Three-Level Cluster-Randomized Experiments in Education. Evaluation Review, 37(6), 445–489. http://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. doi:10.1080/19345741003592428

Konstantopoulos, S. (2009). Incorporating Cost in Power Analysis for Three-Level Cluster-Randomized Designs. Evaluation Review, 33(4), 335–357. doi:10.1177/0193841X09337991

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

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

Reardon, S. F., Shear, B. R., Castellano, K. E., & Ho, A. D. (2015). Using Heteroskedastic Ordered Probit Models to Recover Moments of Coarsened Test Score Distributions. Presented at the National Council for Education Measurement (NCME) Annual Conference.

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

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. doi:10.1177/0193841X14531584

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. doi:10.3102/0162373711423786

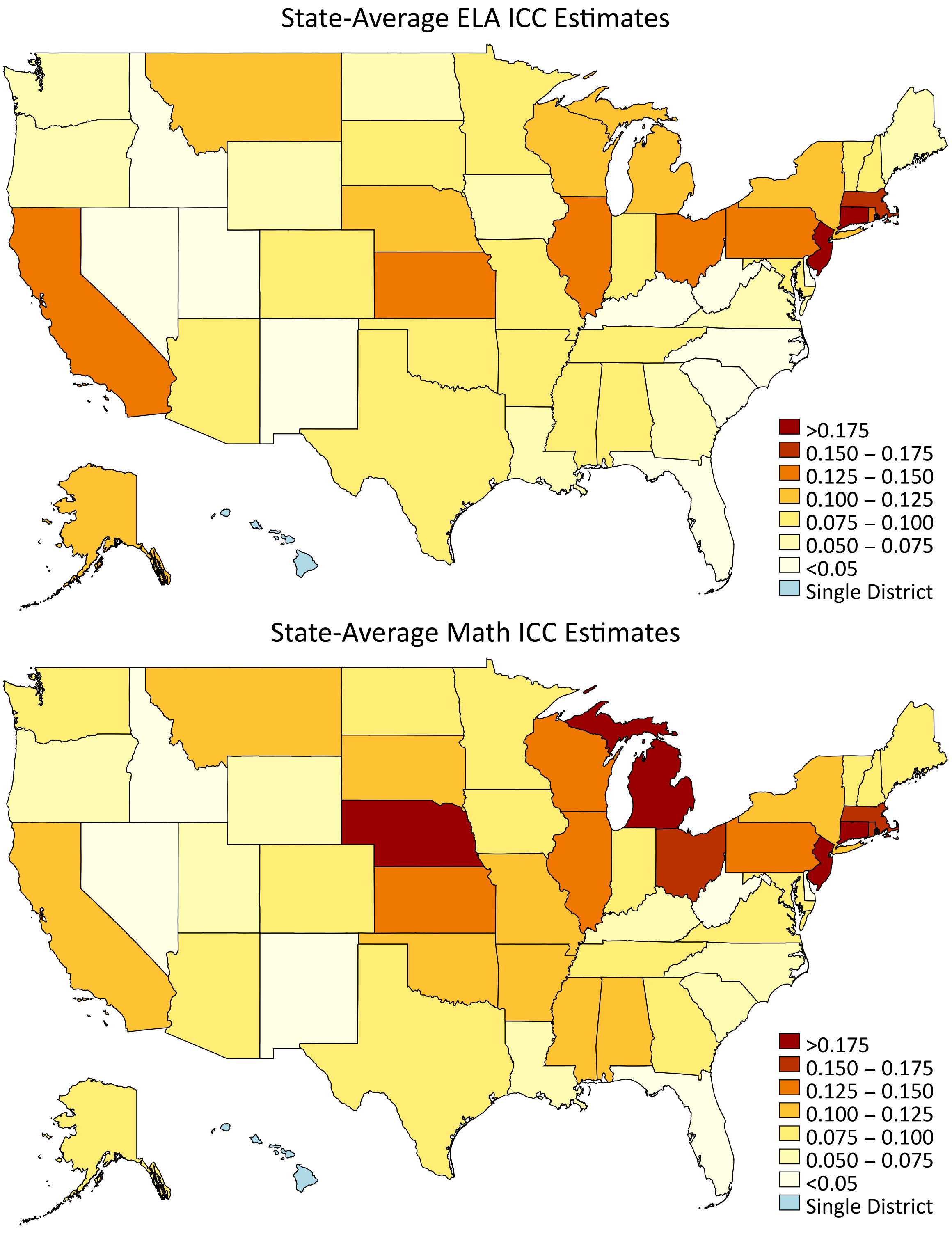
Appendix B. Tables and Figures

Table I: Means and Standard Deviations of ICC Estimates Across States by Subject, Grade and Year

			ELA					Math		
Grade	2009	2010	2011	2012	***Avg***	2009	2010	2011	2012	***Avg***
3	0.080	0.080	0.078	0.082	***0.080***	0.090	0.088	0.089	0.092	***0.09***
	(0.039)	(0.039)	(0.037)	(0.040)	***(0.039)***	(0.042)	(0.042)	(0.041)	(0.044)	***(0.042)***
4	0.082	0.083	0.083	0.085	***0.083***	0.090	0.088	0.091	0.093	***0.091***
	(0.043)	(0.044)	(0.042)	(0.042)	***(0.043)***	(0.044)	(0.045)	(0.044)	(0.046)	***(0.044)***
5	0.084	0.083	0.084	0.086	***0.084***	0.094	0.095	0.096	0.098	***0.096***
	(0.043)	(0.042)	(0.043)	(0.044)	***(0.043)***	(0.048)	(0.046)	(0.046)	(0.046)	***(0.046)***
6	0.090	0.088	0.088	0.092	***0.089***	0.102	0.102	0.100	0.105	***0.102***
	(0.044)	(0.045)	(0.043)	(0.046)	***(0.044)***	(0.048)	(0.049)	(0.045)	(0.046)	***(0.047)***
7	0.088	0.089	0.090	0.089	***0.089***	0.105	0.107	0.107	0.111	***0.108***
	(0.044)	(0.047)	(0.045)	(0.046)	***(0.045)***	(0.048)	(0.049)	(0.049)	(0.050)	***(0.049)***
8	0.088	0.088	0.089	0.090	***0.089***	0.107	0.106	0.108	0.115	***0.109***
	(0.046)	(0.045)	(0.046)	(0.046)	***(0.045)***	(0.050)	(0.046)	(0.050)	(0.053)	***(0.050)***
Avg	***0.085***	***0.085***	***0.085***	***0.087***	***0.086***	***0.098***	***0.098***	***0.099***	***0.102***	***0.099***
	(0.043)	***(0.044)***	***(0.043)***	***(0.044)***	***(0.043)***	***(0.047)***	***(0.047)***	***(0.046)***	***(0.048)***	***(0.047)***

Note: Averages are based on estimated ICCS in each state-grade-year-subject cell. For each state there are approximately 48 observations (4 years x 6 grades x 2 subjects), with the following exceptions: (1) no data was provided for WY in 2010; (2) NE did not use a standard test in ELA in the 2009 test year or in Math in the 2009 or 2010 test years; (3) CA offers multiple test options in grades 7 and 8. No ICCs were computed for these cells.

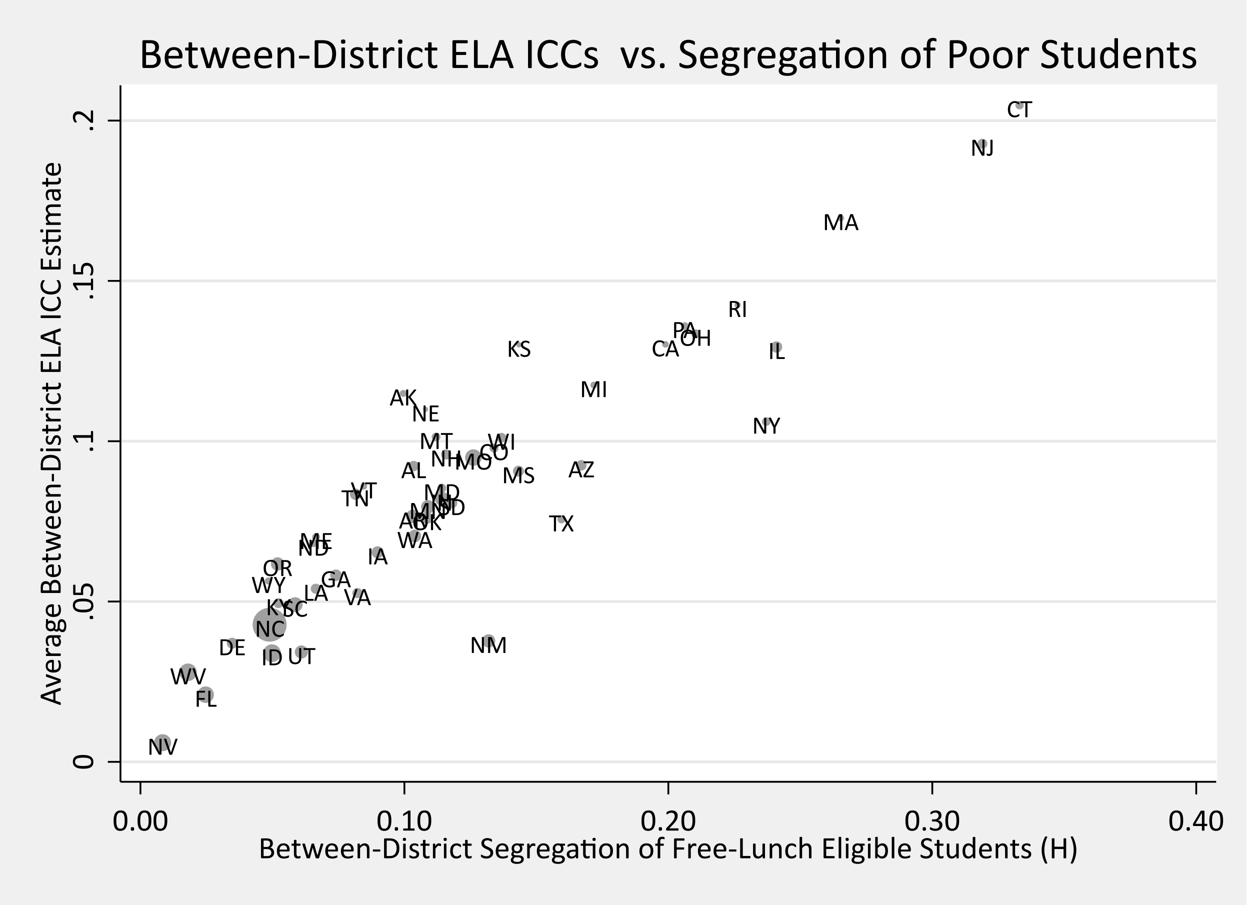
Figure 1: Maps of State Average Between-District ICC Estimates



State-Average ELA ICC Estimates

State-Average Math ICC Estimates

Figure 2: State Between-District ELA ICC Estimates vs. Between-District Segregation in Free-Lunch Eligible Students


Between-District ELA ICCs vs. Segregation of Poor Students
Average Between-District ELA ICC Estimate
Between-District Segregation of Free-Lunch Eligible Students (H)
CT
NJ
MA
RI
PA
OH
CA
IL
KS
MI
AK
NE
NY
MT
WI
NH
MO
CO
AL
MS
AZ
VT
TN
MD
AR
MN
ND
OK
TX
ME
SD
WA
IA
OR
GA
WY
LA
VA
KY
SC
NC
DE
ID
UT
NM
WV
FL
NV