Title Page

Reference Values of Within-district Intraclass Correlations of Academic Achievement by

District Characteristics: Results from a Meta-Analysis of District-specific Values

Running Header: ICC by District Characteristics

E. C. Hedberg

Larry V. Hedges

1

Complete Author Information

E. C. Hedberg, Ph.D (Corresponding Author)

Assistant Professor, Sanford School of Social and Family Dynamics, Arizona State University

Senior Research Scientist at NORC at the University of Chicago

3016 E Wescott Drive

Phoenix, AZ 85050

312 485 5376

ehedberg@asu.edu

Hedberg-Eric@NORC.org

Larry V. Hedges, Ph.D

Board of Trustees Professor at The Institute for Policy Research at Northwestern University

847 491 8899

L-Hedges@Northwestern.edu

Reference Values of Within-district Intraclass Correlations of Academic Achievement by

District Characteristics: Results from a Meta-Analysis of District-specific Values

Abstract

Background

Randomized experiments are often considered the strongest designs to study the impact of educational interventions.   Perhaps the most prevalent class of designs used in large scale education experiments is the cluster randomized design in which entire schools are assigned to treatments. In cluster randomized trials (CRTs) that assign schools to treatments within a set of school districts, the statistical power of the test for treatment effects depends on the within-district school-level intraclass correlation (ICC). Hedges and Hedberg recently computed within-district ICC values in eleven states using three-level models (students in schools in districts) that pooled results across all the districts within each state. While values from these analyses are useful when working with a representative sample of districts, they may be misleading for other samples of districts because the magnitude of ICCs appears to be related to district size. To plan studies with small or nonrepresentative samples of districts, better information are needed about the relation of within-district school-level ICCs to district size.

Objectives

Our objective is to explore the relation between district size and within-district ICCs to provide reference values for math and reading achievement for grades 3-8 by district size, poverty level, and urbanicity level.   These values are not derived from pooling across all districts within a state as in previous work, but are based on the direct calculation of within-district school-level ICCs for each school district.

Research Design

We use mixed models to estimate over 7,000 district-specific ICCs for math and reading achievement in eleven states and for grades 3-8. We then perform a random effects meta-

analysis on the estimated within-district ICCs. Our analysis is performed by grade and subject

for different strata designated by district size (number of schools), urbanicity, and poverty rates.

Introduction

Randomized experiments are often used to evaluate the impact of educational interventions, products, or services. Over the past decade, the number of experiments in education funded by Federal sources has increased considerably (Spybrook & Raudenbush, 2009). Yet, the number of experimental studies reported in the literature has not kept pace (Spybrook, Puente, & Lininger, 2011), possibly due to null findings from studies that are underpowered. The most common experimental designs used in education are cluster randomized trials (CRTs) that assign whole schools to treatments and where the schools are nested within districts.

Cluster randomized trials are commonly used to evaluate interventions in other fields such as health (Hayes, Moulton, & Press, 2009) and have been embraced by the education community. The primary reasons for their adoption is the containment of "spillover" or "contamination" effects (the mixing of treatment and control conditions in a common place) and the efficiencies in delivering place-based services (Bloom, 2005). Blocking (controlling for the effects of schools with similar characteristics) is another common practice in such experiments, with school districts serving as a naturally occurring characteristic on which to block schools. Cluster randomized designs incorporating blocks such as districts are sometimes called multisite cluster randomized trials (MSCRTs).

In multi-level designs the precision of estimates of treatment effects, the statistical power to detect effects, and the minimum effect size that is detectable with a given level of certainty (the minimum detectable effect size) all depend (in part) on the variance decomposition between and within schools (Bloom, 2005; Bloom, Bos, & Lee, 1999; Hedges & Rhoads, 2011; Raudenbush, 1997). In two level designs assigning schools to treatments this variance

6

decomposition is typically summarized by the school-level intraclass correlation coefficient (ICC or ρ), which is the proportion of the total variance that occurs between schools. Therefore planning the sample sizes for a CRT or a MSCRT requires knowledge of the likely value of the school-level intraclass correlation coefficient.

The purpose of this paper is to explore the distribution of within-district ICCs and to provide guidance on ICC values for mathematics and reading achievement across districts of varying size, urbanicity, and levels of poverty. These values will be especially useful to evaluators employing CRTs where schools (clusters) are assigned to treatment condition but are located within a set of districts (blocked sites). The values presented in this paper are unique in that they are not based on three-level (students in schools in districts) mixed models that include entire state data systems (as in Hedges & Hedberg, 2014), but are instead based on school-level ICCs estimated from individual districts. We then summarize these district specific ICCs with a random effects meta-analysis by grade and district subgroups.

Cluster Randomized Trials

The values provided in this paper have specific use for CRTs or MSCRTs where schools are the level of randomization but are blocked by district fixed effects.[1] CRT designs in education are usually three level designs because they involve three-stage sampling where districts (sites) are selected first, then schools (which are statistical clusters), and finally individuals within schools. However when the number of districts is small, they may be considered to have fixed effects since modeling with so few districts would not produce reliable variance components (and thus district effects may be modeled as a set of dummy variables so that the model reduces to two levels of random effects). Thus the model for this design is a two-

---

[1] Of course, this design is only one example of possible three level trials. Another example includes randomizing classrooms within schools.

level model predicting the outcome for the ith student in school j in district k, such as (Spybrook & Raudenbush, 2009)

$$y_{ijk} = \gamma_0 + \gamma_1 W_{jk} + \gamma_2 D_1 + \cdots + \gamma_{K-1} D_{K-1} + r_{jk} + e_{ijk}$$

(1)
$$e_{ijk} \sim N(0, \sigma_W^2), \quad r_{jk} \sim N(0, \sigma_B^2)$$

where γ_0 is the grand mean, γ_1 is the average treatment effect, W_{jk} is a school level variable coded as 0.5 for treatment and -0.5 for control, r_{jk} is the school-level residual with mean 0 and variance σ_B^2, and e_{ijk} is the student level residual with mean 0 and variance σ_W^2.

Given this model, the statistical power of the test for the treatment effect depends on the sample sizes and two other parameters: the within district intraclass correlation (ICC) and the effect size. The within-district school-level ICC is defined as:

$$(2) \qquad \rho = \frac{\sigma_B^2}{\sigma_B^2 + \sigma_W^2}.$$

The effect size (based on the total variation), δ, is defined as

$$(3) \qquad \delta = \frac{\gamma_1}{\sqrt{\sigma_B^2 + \sigma_W^2}}.$$

There are three components to the sample size: the number of districts (sites) selected (K), the number of schools (clusters) per district (J) and the number of individuals within each school (n), which we will assume here are equal in each cluster for simplicity of exposition.

One method to produce a test statistic for testing the null hypothesis of no treatment effect employs the F sampling distribution with 1 and $K(J - 2)$ degrees of freedom. Under the alternative hypothesis, the test statistic has the non-central F distribution has a non-centrality parameter equal to

$$(4) \quad \lambda = \frac{KJ\delta^2}{4\left[\rho + \frac{1-\rho}{n}\right]}.$$

The power of the design is the inverse cumulative (upper tail or survivor) non-central F distribution employing this non-centrality parameter and degrees of freedom $(1, K(J-2))$. For example, the power for a design with effect size 0.2, $n = 20$, $J = 10$, $K = 12$, and an ICC of 0.17 is 0.65.[2]

Finally, it turns out that many different combinations of K, J, and n may give identical (or nearly identical) statistical power. So-called optimal design or optimal allocation methods (which maximize precision or statistical power for a given cost function) are often used to assist in planning cluster randomized designs (see, e.g., Raudenbush, 1997). Optimal allocation depends on cost data but also are a function of the school-level ICC.

In summary, information about within-district school-level intraclass correlations is crucial in planning experiments that use cluster randomized designs either conducted within a single district or using districts as blocks. Intraclass correlations are vital to both estimating the statistical power for a given design, and optimally allocating resources to schools and students. This study adds to the empirical data about such values.

Previous Studies of Design Parameters

Several authors have assembled empirical evidence about intraclass correlations to aide researchers in planning cluster randomized designs. For example, Bloom, Richburg-Hayes, and

---

[2] This calculation can be accomplished in Stata with the following code:
```
. scalar K = 12
. scalar J = 10
. scalar n = 20
. scalar es = .2
. scalar rho = .17
. display  nFtail(1,K*(J-2),(K*J*es^2)/(4*(rho+((1-rho)/n)))),invFtail(1,K*(J-2),.05))
.65473661
```

Black (2007) reported intraclass correlations at several grade levels from five large urban school districts in the Eastern United States that had participated in evaluation studies. (Bloom et al. (2008)) extended the work of Bloom et al. (2007) to provide school-level parameters that extend beyond test scores and include other academic related outcomes in the same five school districts, also providing ICCs for classrooms within schools. (Brandon, Harrison, and Lawton (2013)) in work that provides SAS code for estimating ICCs, also provide upper bound values for Hawaii, a state that is a single school district. Finally, Schochet (2008) provides values for ICCs based on large evaluation studies, but few of these are within-district values.

It is important that the variances of ICC estimates are inversely proportional to the number of schools used and therefore ICC estimates from individual randomized trials (even relatively large ones) are subject to rather large sampling uncertainties (large standard errors). The same thing is true of ICC estimates from all but the largest school districts. Thus the unrepresentative nature of the samples and large sampling uncertainties of estimates given in the studies cited above make them suboptimal as reference values for planning CRTs.

To provide ICC estimates from larger and more representative samples Hedges and Hedberg (2007a) used a set of surveys with large (hundreds to thousands of schools) national probability samples to estimate school-level ICC values for reading and mathematics achievement from Kindergarten through grade 12. ICCs for rural areas were published in Hedges and Hedberg (2007b). Hedges and Hedberg (2011) also provide intraclass correlation estimates by grade, region, and certain school characteristics (such as SES, achievement level, and urbanicity) via the so-called Online Variance Almanac (https://arc.uchicago.edu/reese/variance-almanac-academic-achievement). The ICC estimates are nationally representative and have acceptably small standard errors. However the sampling

designs of the surveys used did not permit the estimation of between-school-district variation. Consequently between-district variation is pooled into between-school variation in the ICC estimates that were computed, which means that the ICCs computed are overestimates of the school-level ICCs (based on three-level models) that are relevant for planning CRTs that use one or a few districts.

To obtain better estimates of within-district school-level ICCs, Hedges and Hedberg (2014) expanded their national database of ICCs by providing values for reading and mathematics achievement based on analyses of State Longitudinal Data Systems (SLDS) in eleven states (Arkansas, Arizona, Colorado, Florida, Louisiana, Kansas, Kentucky, Massachusetts, North Carolina, West Virginia, and Wisconsin; see http://www.ipr.northwestern.edu/research-areas/designparameters/stateva.html). For evaluations across schools where the investigative team is not concerned with school district effects, they provide school-level intraclass correlations based on two-level models that pool district level variation into school level variation. They also provide estimates of the intraclass correlation system from three-level models (i.e. an ICC for district level effects, and another ICC for school level effects). Westine, Spybrook, and Taylor (2014) provide similar values based on SLDS systems for science outcomes.

Why Additional ICC Estimates Are Necessary

The school-level ICC values derived from the statewide three-level models are useful for planning designs that employ a representative sample of districts from a state. However, the research reported in this paper demonstrates that within-district school level ICCs are not constant throughout states, but depend on characteristics of districts, particularly on the number of schools in the district (district size). Therefore, pooled state within-district ICCs may be an

average of dissimilar values that underestimates the ICCs in large districts and overestimates the ICCs in small districts. Moreover, because the pooled state average within-district ICCs give more weight to large districts (because they contribute more information), the pooled state average ICC estimates are particularly poor estimates of the ICCs in smaller school districts. Thus, the estimates of Hedges and Hedberg (2014) based on SLDSs may not be ideal for planning a CRT using a small number of districts, particularly if the districts in the CRT sample are not representative of the state (e.g., if they are smaller districts).

A review of recent published RCTs suggests that the typical RCT uses a small number of districts, usually just one or two. All studies reviewed where the intervention was randomized at the student level used 3 or fewer districts, and half of the studies that randomized at the class or school level also employed 3 or fewer districts. Overall, 66 percent of all studies reviewed used 3 or fewer districts. This is consistent with the idea that most researchers use local education agencies near their institutions to recruit participants and have insufficient resources to manage more than a handful of districts. These results are based on a review of over 20 published articles and reports over the last 3 years, primary from the *American Educational Research Journal, Educational Evaluation and Policy Analysis, Journal of Research on Educational Effectiveness, and The Journal of Experimental Education,* which are primary outlets for education experiments (see Agodini, Harris, Thomas, Murphy, & Gallagher, 2010; Bottge, Grant, Stephens, & Rueda, 2010; Bradshaw, Mitchell, & Leaf, 2010; Calderón, Slavin, & Sánchez, 2011; Fantuzzo, Gadsden, & McDermott, 2011; Fulmer & Frijters, 2011; Gersten, Dimino, Jayanthi, Kim, & Santoro, 2010; Goodson et al., 2011; Hamre et al., 2012; Isenberg et al., 2009; Kim, Capotosto, Hartry, & Fitzgerald, 2011; Lane et al., 2011; Laura, McMeeking, Orsi, & Cobb, 2012; Marley, Levin, & Glenberg, 2010; Marley, Szabo, Levin, & Glenberg,

2011; McQuillin, Smith, & Strait, 2011; Olson et al., 2012; Phelan, Choi, Vendlinski, Baker, & Herman, 2011; Reis, McCoach, Little, Muller, & Kaniskan, 2011; Rose, Woolley, Orthner, Akos, & Jones-Sanpei, 2012; Sarama, Clements, Wolfe, & Spitler, 2012; Slavin, Cheung, Holmes, Madden, & Chamberlain, 2012; Springer et al., 2012; VanDerHeyden, McLaughlin, Algina, & Snyder, 2012; Vaughn, Klingner, et al., 2011; Vaughn, Wexler, et al., 2011; Wirkala & Kuhn, 2011; Wolf et al., 2010).

Analysis Plan

The purpose of our analysis is to estimate typical within-district ICCs by subject, grade, district size, urbanicity, and poverty status. Our analysis follows three steps. First, specific to subject and grade, we estimate district-specific school-level ICCs using 11 state data systems: Arkansas, Arizona, Colorado, Florida, Kansas, Kentucky, Louisiana, Massachusetts, North Carolina, West Virginia, and Wisconsin. All data was from the 2009-2010 school year with the exception of Florida, which supplied data from the 2006-2007 school year, Louisiana, which supplied data from the 2012-2013 school year, and West Virginia, which supplied data from the 2011-2012 school year. Since all states test in grades $3 - 8$, we focused our analysis only on these grades.

Whether a district was included in the analysis was evaluated separately for each grade. Eligible districts were those that had test scores in at least 2 schools that served a particular grade (since ICCs are undefined in a district with a single school) and had a harmonic mean number of at least 2 student scores per school. We use the harmonic mean since it is less prone to outliers. We used a threshold of 2 students because the variance of the ICC, given below in (8), increases exponentially for harmonic means of fewer than 2 students (regardless of the value of the ICC).

13

Each state employed a different achievement test, namely the Augmented Benchmark Examination (Arkansas), Arizona's Instrument to Measure Standards, Colorado's Student Assessment Program the Florida Comprehensive Assessment Test, the Kansas Assessment Program, the Commonwealth Accountability Testing System (Kentucky), Louisiana's Integrated Educational Assessment Program, Massachusetts Comprehensive Assessment System, the North Carolina End of Grade Tests, West Virginia's WESTEST, and the Wisconsin Knowledge and Concepts Examination.

Second, we compiled our district-specific ICCs into a database and assigned subgroup identifiers. Employing the CCD (Keaton, 2012) we estimated the 10th, 25th, and 50th percentiles of district size that serve students in grades 3, 4, 5, 6, 7, and 8. Our percentile analysis used the student count to weight the school records. Thus, we found the district size percentiles from the student point of view. The 10th percentile means that 10 percent of student are served by a districts of a particular size. Weighting the districts by students served by grade, we found for grades 3 and 4 that the 10th percentile of district size was 3 schools, the 25th percentile was 5 schools, and the 50th percentile was 10 schools. In grades 5 and 6, the 10th, 25th, and 50th percentiles were 2, 5, and 11 schools, respectively. In grade 7, the 10th, 25th, and 50th percentiles were 2, 3, and 6 schools, respectively. Finally, in grade 8, the 10th, 25th and 50th percentiles were 2, 3, and 7 schools, respectively. The sample of district specific ICCs was then divided into four groups for each grade, using the 10th, 25th, and 50th percentiles of district size as cut points. These size grouping are noted as "very small," "small," "medium," and "large" districts. We include these school sizes in the results tables for clarity.

Finally, we summarize the district-specific ICCs using a random effects meta-analytic approach (Borenstein, Hedges, Higgins, & Rothstein, 2011; Hedges & Vevea, 1998) as detailed

below. We do this by grade and subject and also for district size, poverty status, and urbanicity groups. Poverty is defined as a two-group variable indicating either that the district has a) fewer than 50 percent of its students eligible for free or reduced price lunch or b) 50 percent or more of students are eligible for free or reduced priced lunch. Urbanicity is also a two-group variable indicating either that the district is a) primarily not in an urban area or b) primarily in an urban area. Urban areas are defined by Common Core of Data standards. Urban areas meet of the following criteria: it is a territory inside an urbanized area and inside a principal city with population of 250,000 or more, a territory inside an urbanized area and inside a principal city with population less than 250,000 and greater than or equal to 100,000, or a territory inside an urbanized area and inside a principal city with population less than 100,000.

We provide ICCs by district size categories because there is a relationship between the log number of schools and the value of the ICC (presented in the results section). In addition to district size, many studies are focused on impoverished populations and/or urban populations, which may have different ICC values. To that end, we also provide results by district size for districts with more and fewer than 50 percent of students eligible for free or reduced price lunch and for districts are or are not located in urban areas. For example, researchers conducting evaluation studies in large urban school districts will find Tables 6 and 7 most useful.

Statistical Methodology

Estimating District Specific ICCs

The district-specific ICCs were estimated by selecting each eligible district in each state, selecting all students within a specific grade, and setting the outcome to either the reading or math score. Once selected, we estimated an unconditional two-level mixed model using restricted maximum likelihood,

$$(5) \qquad y_{ij} = \mu + \eta_j + \varepsilon_{ij},$$

where y_{ij} is the score from the *ith* student from school *j*, μ is the average of school average, η_j is the school random effect, and ε_{ij} is the within-school student random effect. The within-school variance component (the variance of the ε_{ij}'s) is σ_W^2 and the between-school variance component (the variane of the η_j's) is σ_B^2. The estimated intraclass correlation is obtained using the estimated variance components as specified in (2).

Random Effects Meta-Analysis of District-Specific ICCs

Our analysis produced several thousand district-specific ICCs, many of which are estimated from small districts where concerns about privacy are relevant. Moreover, there is considerable variation in estimates from similar districts, undoubtedly due to random sampling error. Therefore, instead of providing tables with several hundreds of estimates, we instead summarize our results by presenting average ICCs derived from a random effects meta-analysis (Hedges & Vevea, 1998). Below we provide a brief overview of this procedure in the context of our study.

The goal of a meta-analysis is to summarize the results of a series of estimations in order to provide guidance on the expected "effect." In our case the effect is the ICC, and we wish to estimate the population's typical ICC, based on a given set of *k* estimates, for use in planning CRTs. If we assumed that the true ICC was the same in all districts (in other words treating the districts as fixed effects), we would conceptualize any estimate, Y_i, as the sum of the true effect, θ, and the sampling error, ε_i

$$(6) \qquad Y_i = \theta + \varepsilon_i.$$

Of course, we don't know the true effect, only the estimates and the sampling variation associated with them. We can achieve an estimate of the true effect by using the inverse variance of the estimate as a weighting variable. For our ICCs, the estimated variance for the ith ICC is (Fisher, 1925):

$$(7) \quad V_i = \frac{2(1-\rho_i)^2(1+(n_i-1)\rho_i)^2}{n_i(n_i-1)(m_i-1)},$$

where n_i is the harmonic mean number of students per school in the district and m_i is the number of schools in that district. The weight for each ICC is then simply

$$(8) \quad W_i = V_i^{-1},$$

and the estimate of the true ICC would be defined as

$$(9) \quad \hat{\mu}_\rho = \frac{\sum W_i Y_i}{\sum W_i}.$$

However, a weakness of this approach is that we cannot assume the same ICC for all districts, even in a subgroup, for two reasons. First, we are making a generalization beyond the observed results. This introduces a random effect beyond the sampling error that must be addressed. A second, more nuanced, set of problems with the fixed effects approach is that each state employs a different standardized test (at least in our data), each state organizes districts in a slightly different way, and the way districts organize their students is not universal. Thus ICCs are derived from slightly different processes across our observed districts. As a result, we must employ a random effects approach to the meta-analysis.

In a random effects meta-analysis, we conceptualize the estimate, Y_i, as the sum of the average of the true effects, μ_θ, the district's deviance from the average of the true effect, ζ_i, and the sampling error, ε_i:

$$(10) \quad Y_i = \mu_\theta + \zeta_i + \varepsilon_i .$$

We must therefore account for the variance associated with both sampling errors and the variance of the true district values around the average true effect. This is accomplished by estimating the between-district variance of the ICCs, τ^2. This quantity can be estimated with a method of moments estimator, given in <u>Hedges and Vevea (1998)</u> as

$$(11) \quad \hat{\tau}^2 = \frac{Q - k - 1}{\sum W_i - \sum W_i^2 / \sum W_i} ,$$

where

$$(12) \quad Q = \sum W_i Y_i^2 - \frac{\left(\sum W_i Y_i\right)^2}{\sum W_i} .$$

When the expression in (11) is non-negative and $\tau^2 = 0$ if the expression given in (11) is negative. Also note that this estimation makes no assumptions about the underlying distribution of the effects (i.e., ICCs). Therefore, it is still an unbiased estimate of the variation in ICCs. However, we do not recommend the use of this variance component to compute a range a plausible ICC values (e.g., $\hat{\mu}_\rho \pm \hat{\tau} \times z_{\alpha/2}$) because the distribution is not normal.

To test the null hypothesis that $\tau = 0$, we use the fact that Q, given in (12), has a χ^2 distribution with k-1 degrees of freedom when $\tau = 0$. With this estimate, we calculate the random effects weight for each ICC

$$(13) \quad W_i^* = \left(V_i + \tau^2\right)^{-1} .$$

The summary reported in our results is then the weighted mean of the observed ICCs

$$(14) \quad \hat{\mu}_\rho^* = \frac{\sum W_i^* Y_i}{\sum W_i^*} ,$$

and its standard error (the square root of the inverse of the sum of the weights),

$$(15) \quad SE\left(\widehat{\mu}_\rho^*\right) = \sqrt{\left(\sum W_i^*\right)^{-1}}.$$

Note that if τ is estimated to be negative, in which case we truncate $\widehat{\tau}$ to 0, the weights are simply the fixed effect weights as in (8) and the random effects analysis is identical to the fixed effects analysis.

Database of District-Specific Estimates

This section describes the database of district-specific estimates that we compiled. In a small number of cases, the ICC estimates were quite large and could inflate the estimate of τ^2. Therefore, to avoid allowing outliers to have disproportionate influence on our estimate of τ^2, we removed the top one percent of estimates, redacting 71 estimates from our input data greater than the 99th percentile of the estimates (0.557). While this did not have a measurable impact on the mean estimate, it substantially decreased the estimate of τ^2. This resulted in a set of 3,555 ICCs for mathematics achievement and 3,557 ICCs for reading achievement. Table 1 presents the number of eligible districts by state, grade, and subject. Table 1 also includes the number of students used in the estimates.

TABLE 1 HERE

The results presented in this paper are based on over 3.1 million students. Of the ICCs computed, 16 percent are from urban areas and about 58 percent are from high poverty areas. About 57 percent of the non-urban areas and 62 percent of the urban areas are high poverty. Figure 3 presents the number of ICCs estimated by district size, urbanicity, and poverty. The modal ICC is estimated from a non-urban, high poverty, medium sized district, followed by a

similar small district. The next most common ICC is estimated from a non-urban, low poverty,

small district. Larger districts were more prevalent in urban areas, as would be expected.

The mean ICC estimated from all districts, grades, and subjects was 0.094, with a

standard deviation of 0.092. The distribution is highly skewed, with a median of 0.056. The

estimated ICCs for math had a mean of 0.104 and standard deviation of 0.104. The 10th, 25th,

50th, 75th and 90th percentiles for math were 0.001, 0.027, 0.076, 0.149, and 0.246, respectively.

The estimated ICCs for reading were generally lower with a mean of 0.084 and standard

deviation of 0.092. The 10th, 25th, 50th, 75th and 90th percentiles for reading were <0.001, 0.018,

0.056, 0.118, and 0.200, respectively.

Figure 4 presents boxplots of the estimated district-specific ICCs by subject, grade, and

district size. Each boxplot presents a highly skewed distribution. In all districts, math ICCs tend

to have a larger median than the reading ICCs, and the medians generally rise with grade level.

The variance also increases with grade levels. Examining the boxplots for the very small

districts, those below the 10th percentile, we see the reverse pattern: ICCs and the variance

decrease with grade. The small and medium school districts do not display a consistent pattern

with grades, except that the 8th grade variance is larger. Finally, large school districts are more

reflective of the overall pattern.

Finally, as support for presenting meta-analyses by district size, we estimated unweighted

correlations between the district-specific ICC by the log of district size (number of schools) for

each subject and grade. The correlation coefficients ranged from 0.52 to 0.75 with a median of

0.70, which supports the claim that ICCs are related to district size.

Results of Meta-Analyses

Tables 2-11 present the estimated mean ICCs and τ for math and reading achievement by grade and district size. In these tables we also present the empirical 25th, 50th (the median), and 75th percentiles to give a sense of the observed distribution and variance. Each table is organized in a series of horizontal panels, each for a district size category, with rows for each grade. The number of districts used for the analysis is denoted as k. Tables 2 and 3 present results for all districts. These results are useful for research designs that sample districts with a variety of characteristics and are not limited to only impoverished or rural areas.

However, other designs may be more specific. To serve those researchers, Tables 4 and 5 present results for non-urban districts. Tables 6 and 7 present results for urban districts. Tables 8 and 9 present results for low-poverty (less than 50 percent of students who are eligible for free or reduced price lunch) districts. Finally, Tables 10 and 11 present results for high-poverty (at least 50 percent of students who are eligible for free or reduced price lunch) districts. In this section we present some patterns (or lack of patterns) of interest. Overall, each pattern noted here will have exceptions, but the following will provide some basic insight into the distribution of ICCs.

<div align="center">

TABLE 2 HERE

TABLE 3 HERE

</div>

Results for all districts and comparison with statewide estimates

As is typical in other studies, we generally see that ICCs for mathematics are higher than ICCs estimated for reading. This is a relatively stable pattern, but there are exceptions. In medium-sized districts, the reading ICCs are larger for grades 4 and 6. They are also larger in small 8th grade districts. For mathematics achievement, the average within-district ICC derived

<div align="center">

21

</div>

from individual three level models (students nested within schools nested within districts) from each state is about 0.11 for grades 3, 4, and 5. The meta-analysis results in smaller ICCs for grades 3, 4, and 5: 0.07, 0.06, and 0.08, respectively. While our 7th grade estimate is also smaller, 0.10 vs. 0.13, our estimates for 8th grade are larger than the three level models: 0.17 vs. 0.16.

We also observe smaller estimated mean ICCs in our analysis for reading achievement in all grades, compared to the results from analyses that pool all data across districts within each state. In grades 3 and 4, the average results from the state data that pool the information across districts are much larger than the estimated average of the district-specific estimates from the meta-analysis: 0.10 versus 0.05 and 0.10 versus 0.06, respectively. In grade 7, the result from our meta-analysis is smaller than the average of the district-specific estimates, 0.07 versus 0.10.

To contextualize the results presented in this section to results from estimates that pool all data across districts, we provide the following guidance. The results in this study are appropriate for planning targeted samples that do not represent an entire state. Conversely, the results from data that pool information across all districts are meant to inform designs that draw a sample from all districts.

Results by district size

In most grades, the ICCs are larger in larger districts. For example, in Table 2, the grade 3 math ICCs for very small, small, medium, and large districts are 0.009, .012, 0.084, and 0.118, respectively. This general pattern holds for all grades in math except grade 7, where the small districts have a larger ICC than the medium districts. Another notable feature is that the pattern is not exceptionally linear, with the larger districts having much larger ICCs than the smaller districts.

Also of note is that the largest districts show similar ICCs for earlier grades compared with the ICCs from three level models. For example, the mean math ICC from the meta-analysis for grade 3 in the largest districts is 0.118, which is similar to the average from the three-level model analyses (0.112). This supports the hypothesis that the three level models are unduly influenced by larger districts.

Results by grade

Previous investigations found that ICCs generally increase with grade level (Hedges and Hedberg 2007a, 2007b, and 2014). While we again find this is true in broad strokes, closer examination reveals a more complicated picture. In all districts, we observe a pattern for math in which mean ICCs increase, with grades 3, 4, 5, 6, 7, and 8 having mean ICCs of 0.072, 0.063, 0.080, 0.084, 0.097, and 0.169, respectively. This pattern is not evident in reading, with the ICCs appearing to "bounce around" for grades 3-7. These patterns hold in the smaller districts as well, although there is a linear increase of the ICCs by grade in the largest districts.

Results by urbanicity

Tables 4 and 5 present ICCs for mathematics and reading achievement for non-urban districts, while Tables 6 and 7 present math and reading ICCs for urban districts. Some combinations of district size and urbanicity were not represented in our data and thus meta-analysis was not possible. For districts of any size, we generally find ICCs in urban areas are larger for the lower grades than those in non-urban areas. In the higher grades, the non-urban areas tend to have larger ICCs, especially in 8th grade.

<div align="center">TABLE 4 HERE</div>

<div align="center">TABLE 5 HERE</div>

<div align="center">TABLE 6 HERE</div>

TABLE 7 HERE

Results by poverty

Tables 8 and 9 present ICCs for mathematics and reading achievement for districts with less than a 50 percent rate of free or reduced price lunch eligible students, while Tables 10 and 11 present math and reading ICCs for districts with at least half of students eligible for free or reduced price lunch. For districts of any size, we generally find that ICCs in "high poverty" districts are larger than those in "low poverty" districts. The exception to this pattern is 7th grade reading, where the high poverty ICC is slightly lower than the low poverty ICC.

In the smaller districts, we generally see that the high poverty ICCs are higher than the low poverty ICCs in the earlier grades (3-5). In grades 6-8, however, it is the low poverty ICCs that are smaller in the smaller districts. In the medium size districts, the reading ICCs are lower in most grades for low poverty compared to high poverty districts, whereas the math ICCs do not seem to follow a pattern. Finally, in the largest districts, the math ICCs are higher in the high poverty districts for the lower grades, and are generally higher for reading in most grades except 8th.

TABLE 8 HERE

TABLE 9 HERE

TABLE 10 HERE

TABLE 11 HERE

Variation in estimated ICCs

In Tables 2 – 11 we also report the variation in ICCs for grade and district size combinations as the standard deviation $\hat{\tau}$. We tested this estimate against the null hypothesis that $\tau = 0$ and marked estimates with less than a 5 percent chance of Type I error in rejecting the

null hypothesis that $\tau = 0$; in other words, we follow standard practice and mark statistically significant variance.   For results of all districts (that is, ignoring district size), we universally found significant variation in ICCs. This held consistently for the largest districts, although we found less consistent evidence of variation in ICCs for the smaller districts. In many cases, our estimate of τ was negative and truncated at 0. In such cases, we entered the letter "a" into the table.

Discussion

In this study we provide expected ICCs by grade and subject in a variety of contexts that will also be of interest to evaluation researchers, namely district size, urbanicity, and poverty status. While we generally found expected patterns, the smaller districts presented a picture that was less consistent. Perhaps the sampling variability associated with smaller districts creates difficulty in uncovering patterns of results, or perhaps such patterns do not exist.

To our knowledge, this is the first investigation of the distribution of within-district ICCs. One of the more important findings of this study is simply that these ICCs tend to be quite small for earlier grades in the smaller districts. This is particularly important for planning interventions in these settings, because pretests on academic achievement that might be used as covariates to improve statistical power are less frequently available in administrative data on younger children. Given the small ICC estimates, the practice of spending resources of pretests may not be necessary.

Finally, it is worth noting that the more heterogeneous districts, in terms of expected ICCs, are the largest districts. This is not surprising given the diversity found in large urban areas. It does, however, highlight the need to utilize local data when available. Although previous publications have provided such data from the handful of large urban districts that have

been studied in previous evaluations, our results provide some guidance in other situations where little data are available.

Limitations

We have two limitations of this study to outline. The first limitation was the amount of data available to produce ICC estimates to analyze. To be sure, we employ a relatively large amount of data compared to most education studies, but for this paper out unit of analysis is the district. With 11 states, we only have between 400 and 780 districts per grade to analyze. Thus, more detailed breakdowns by urbanicity and poverty produce too few estimates in certain cells to produce reliable means. Until more data are available, the conservative approach would be to employ the larger of available ICCs for more targeted sample (e.g., impoverished urban areas). Finally, we would also like to have a more detailed urbanicity breakdown, but again we are restricted by our sample size.

Conclusions

We have presented empirical evidence about design parameters useful in planning CRT experiments that used academic achievement as outcomes and where districts of a particular size or type are employed. Our estimates are means derived from random effects meta-analyses and are presented along with standard errors that provide some sense of the sampling error inherent in these estimates. We now turn to the question of how best to use these design parameters in planning CRTs and what some limitations might be.

The intraclass correlation values reported in this tabulation differ from the national values reported by Hedges and Hedberg (2007a) and their more recent work (2014). While the evidence reported in this paper is based on near-census data from several state longitudinal data systems, it is data from only eleven states and only in grades 3 to 8. While our estimates should

be more relevant for some studies than national estimates like those of Hedges and Hedberg (2007a), there are significant heterogeneities across large districts. However, for certain applications, the results in this paper may prove more useful than estimates derived from models that pool results across states. This study has also revealed that the distribution of within-district school-level ICCs is highly skewed, with numerous very small ICCs and a small number of large values. However, meta-analyses reveal that small districts have ICCs of relatively uniform size, with measures of variation seldom statistically differing from zero. A final caution is that the estimates reported here are based on state assessments and thus would be less relevant to studies using achievement tests that are not aligned with instruction.

Example Power Analyses for CRTs

Putting such limitations aside, these values can be used with several pieces of software designed for multi-level power computations, including *Optimal Design* for Windows (Spybrook, Raudenbush, Liu, Congdon, & Martínez, 2006), *RDPOWER* for Stata (Hedberg, 2012), and commercial software such as *CRT-Power* (Borenstein, Hedges, & Rothstein, 2012). Here, we provide an example with immediate commands in Stata.

Suppose we were to perform an experiment that impacts mathematics achievement of third graders in eight large school districts, with each district having 12 schools. We plan to collect data on 30 students in each of these 96 schools. Power to detect an effect size of 0.2 is computed using the noncentrality parameter (4) by entering the following into Stata

```
. scalar K = 12
. scalar J = 8
. scalar n = 30
. scalar es = .2
. scalar rho = .118
. display  nFtail(1,K*(J-2),(K*J*es^2)/(4*(rho+((1-rho)/n)))),invFtail(1,K*(J-2),.05))
```

The result gives the statistical power of a two-tailed test at the 0.05 level of significance as 0.71.

Summary

        The main finding of this study is that district size matters. In some cases, employing smaller districts (and thus fewer schools) may yield better power because the ICCs are that much smaller. While which districts participate in a study is rarely under investigator control, suppose we are designing a study for third grade math, and the choice is between using 4 medium districts with 10 schools each (ICC = 0.031) and 4 large districts with 20 schools each (ICC = 0.118). Holding other factors constant, and assuming a fixed effects design, the smaller districts yield slightly more power for each effect size. Of course, employing a single large district, even with a larger ICC, may have the practical benefit of only having to recruit a single education agency.

References

Agodini, R., Harris, B., Thomas, M., Murphy, R., & Gallagher, L. (2010). Achievement Effects of Four Early Elementary School Math Curricula: Findings for First and Second Graders. NCEE 2011-4001. *National Center for Education Evaluation and Regional Assistance*.

Bloom, H. S. (2005). *Learning more from social experiments: Evolving analytic approaches*: Russell Sage Foundation.

Bloom, H. S., Bos, J. M., & Lee, S.-W. (1999). Using Cluster Random Assignment to Measure Program Impacts Statistical Implications for the Evaluation of Education Programs. *Evaluation Review, 23*(4), 445-469.

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.

Bloom, H. S., Zhu, P., Jacob, R., Raudenbush, S., Martinez, A., & Lin, F. (2008). *Empirical issues in the design of group-randomized studies to measure the effects of interventions for children*: MDRC New York.

Borenstein, M., Hedges, L., & Rothstein, H. (2012). CRT-Power *Teaneck, NJ: Biostat*.

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2011). *Introduction to meta-analysis*: Wiley. com.

Bottge, B. A., Grant, T. S., Stephens, A. C., & Rueda, E. (2010). Advancing the math skills of middle school students in technology education classrooms. *NASSP Bulletin, 94*(2), 81-106.

Bradshaw, C. P., Mitchell, M. M., & Leaf, P. J. (2010). Examining the effects of schoolwide positive behavioral interventions and supports on student outcomes results from a

randomized controlled effectiveness trial in elementary schools. *Journal of Positive Behavior Interventions, 12*(3), 133-148.

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.

Calderón, M., Slavin, R., & Sánchez, M. (2011). Effective instruction for English learners. *The Future of Children, 21*(1), 103-127.

Fantuzzo, J. W., Gadsden, V. L., & McDermott, P. A. (2011). An integrated curriculum to improve mathematics, language, and literacy for Head Start children. *American Educational Research Journal, 48*(3), 763-793.

Fisher, R. A. (1925). *Statistical methods for research workers*: Genesis Publishing.

Fulmer, S. M., & Frijters, J. C. (2011). Motivation During an Excessively Challenging Reading Task: The Buffering Role of Relative Topic Interest. *The Journal of Experimental Education, 79*(2), 185-208.

Gersten, R., Dimino, J., Jayanthi, M., Kim, J. S., & Santoro, L. E. (2010). Teacher Study Group Impact of the Professional Development Model on Reading Instruction and Student Outcomes in First Grade Classrooms. *American Educational Research Journal, 47*(3), 694-739.

Goodson, B., Wolf, A., Bell, S., Turner, H., Finney, P. B., & Garcia, S. (2011). Effectiveness of a Program to Accelerate Vocabulary Development in Kindergarten (VOCAB): First Grade Follow-up Impact Report and Exploratory Analyses of Kindergarten Impacts.

Hamre, B. K., Pianta, R. C., Burchinal, M., Field, S., LoCasale-Crouch, J., Downer, J. T., . . . Scott-Little, C. (2012). A Course on Effective Teacher-Child Interactions Effects on

Teacher Beliefs, Knowledge, and Observed Practice. *American Educational Research Journal, 49*(1), 88-123.

Hayes, R. J., Moulton, L. H., & Press, C. (2009). *Cluster randomised trials*: CRC press London.

Hedberg, E. C. (2012). RDPOWER: Stata module to perform power calculations for random designs. *Statistical Software Components*.

Hedges, L. V., & Hedberg, E. C. (2007a). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis, 29*(1), 60-87.

Hedges, L. V., & Hedberg, E. C. (2007b). Intraclass correlations for planning group randomized experiments in rural education. *Journal of Research in Rural Education, 22*(10), 22-10.

Hedges, L. V., & Hedberg, E. C. (2011). Variance Almanac of Academic Achievement. 2013, from https://arc.uchicago.edu/reese/variance-almanac-academic-achievement

Hedges, L. V., & Hedberg, E. C. (2014). *Intraclass Correlations and Covariate Outcome Correlations for Planning 2 and 3 Level Cluster Randomized Experiments in Education*.

Hedges, L. V., & Rhoads, C. H. (2011). Correcting an analysis of variance for clustering. *British Journal of Mathematical and Statistical Psychology, 64*(1), 20-37.

Hedges, L. V., & Vevea, J. L. (1998). Fixed-and random-effects models in meta-analysis. *Psychological Methods, 3*(4), 486.

Isenberg, E., Glazerman, S., Bleeker, M., Johnson, A., Lugo-Gil, J., Grider, M., . . . Britton, E. (2009). Impacts of Comprehensive Teacher Induction: Results from the Second Year of a Randomized Controlled Study. NCEE 2009-4072. *National Center for Education Evaluation and Regional Assistance*.

Keaton, P. (2012). *Documentation to the NCES Common Core of Data Public Elementary/Secondary School Universe Survey: School Year 2010–11 (NCES 2012-338rev).*

Kim, J. S., Capotosto, L., Hartry, A., & Fitzgerald, R. (2011). Can a mixed-method literacy intervention improve the reading achievement of low-performing elementary school students in an after-school program? Results from a randomized controlled trial of READ 180 enterprise. *Educational Evaluation and Policy Analysis, 33*(2), 183-201.

Lane, K. L., Harris, K., Graham, S., Driscoll, S., Sandmel, K., Morphy, P., . . . Schatschneider, C. (2011). Self-Regulated Strategy Development at Tier 2 for Second-Grade Students With Writing and Behavioral Difficulties: A Randomized Controlled Trial. *Journal of Research on Educational Effectiveness, 4*(4), 322-353.

Laura, B., McMeeking, S., Orsi, R., & Cobb, R. B. (2012). Effects of a teacher professional development program on the mathematics achievement of middle school students. *Journal for Research in Mathematics Education, 43*(2), 159-181.

Marley, S. C., Levin, J. R., & Glenberg, A. M. (2010). What Cognitive Benefits Does an Activity-Based Reading Strategy Afford Young Native American Readers? *The Journal of Experimental Education, 78*(3), 395-417.

Marley, S. C., Szabo, Z., Levin, J. R., & Glenberg, A. M. (2011). Investigation of an activity-based text-processing strategy in mixed-age child dyads. *The Journal of Experimental Education, 79*(3), 340-360.

McQuillin, S., Smith, B., & Strait, G. (2011). Randomized evaluation of a single semester transitional mentoring program for first year middle school students: a cautionary result

for brief, school‐based mentoring programs. *Journal of Community Psychology, 39*(7), 844-859.

Olson, C. B., Kim, J. S., Scarcella, R., Kramer, J., Pearson, M., van Dyk, D. A., . . . Land, R. E. (2012). Enhancing the Interpretive Reading and Analytical Writing of Mainstreamed English Learners in Secondary School Results From a Randomized Field Trial Using a Cognitive Strategies Approach. *American Educational Research Journal, 49*(2), 323-355.

Phelan, J., Choi, K., Vendlinski, T., Baker, E., & Herman, J. (2011). Differential Improvement in Student Understanding of Mathematical Principles Following Formative Assessment Intervention. *The Journal of Educational Research, 104*(5), 330-339.

Raudenbush, S. W. (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods, 2*(2), 173-185.

Reis, S. M., McCoach, D. B., Little, C. A., Muller, L. M., & Kaniskan, R. B. (2011). The effects of differentiated instruction and enrichment pedagogy on reading achievement in five elementary schools. *American Educational Research Journal, 48*(2), 462-501.

Rose, R. A., Woolley, M. E., Orthner, D. K., Akos, P. T., & Jones-Sanpei, H. A. (2012). Increasing Teacher Use of Career-Relevant Instruction A Randomized Control Trial of CareerStart. *Educational Evaluation and Policy Analysis, 34*(3), 295-312.

Sarama, J., Clements, D. H., Wolfe, C. B., & Spitler, M. E. (2012). Longitudinal evaluation of a scale-up model for teaching mathematics with trajectories and technologies. *Journal of Research on Educational Effectiveness, 5*(2), 105-135.

Schochet, P. Z. (2008). Statistical power for random assignment evaluations of education programs. *Journal of Educational and Behavioral Statistics, 33*(1), 62-87.

Slavin, R. E., Cheung, A., Holmes, G., Madden, N. A., & Chamberlain, A. (2012). Effects of a Data-Driven District Reform Model on State Assessment Outcomes. *American Educational Research Journal*.

Springer, M. G., Pane, J. F., Le, V.-N., McCaffrey, D. F., Burns, S. F., Hamilton, L. S., & Stecher, B. (2012). Team Pay for Performance Experimental Evidence From the Round Rock Pilot Project on Team Incentives. *Educational Evaluation and Policy Analysis, 34*(4), 367-390.

Spybrook, J., Puente, A. C., & Lininger, M. (2011). An examination of the impact of changes in federal policies on the landscape of educational research in the USA. *Effective Education, 3*(2), 83-88.

Spybrook, J., & Raudenbush, S. W. (2009). An examination of the precision and technical accuracy of the first wave of group-randomized trials funded by the Institute of Education Sciences. *Educational Evaluation and Policy Analysis, 31*(3), 298-318.

Spybrook, J., Raudenbush, S. W., Liu, X.-f., Congdon, R., & Martínez, A. (2006). Optimal design for longitudinal and multilevel research: Documentation for the "Optimal Design" software. *Survey Research Center of the Institute of Social Research at University of Michigan*.

VanDerHeyden, A., McLaughlin, T., Algina, J., & Snyder, P. (2012). Randomized Evaluation of a Supplemental Grade-Wide Mathematics Intervention. *American Educational Research Journal, 49*(6), 1251-1284.

Vaughn, S., Klingner, J. K., Swanson, E. A., Boardman, A. G., Roberts, G., Mohammed, S. S., & Stillman-Spisak, S. J. (2011). Efficacy of Collaborative Strategic Reading With Middle School Students*. *American Educational Research Journal, 48*(4), 938-964.

Vaughn, S., Wexler, J., Roberts, G., Barth, A. A., Cirino, P. T., Romain, M. A., . . . Denton, C.
A. (2011). Effects of individualized and standardized interventions on middle school
students with reading disabilities. *Exceptional children, 77*(4), 391-407.

Westine, C. D., Spybrook, J., & Taylor, J. A. (2014). An Empirical Investigation of Variance
Design Parameters for Planning Cluster-Randomized Trials of Science Achievement.
Evaluation Review, 0193841X14531584.

Wirkala, C., & Kuhn, D. (2011). Problem-Based Learning in K–12 Education Is it Effective and
How Does it Achieve its Effects? *American Educational Research Journal, 48*(5), 1157-
1186.

Wolf, P., Gutmann, B., Puma, M., Kisida, B., Rizzo, L., Eissa, N., & Carr, M. (2010). Evaluation
of the DC Opportunity Scholarship Program: Final Report. NCEE 2010-4018. *National
Center for Education Evaluation and Regional Assistance*.

Figure 1 Number of Estimates by Grade-specific District Size, Urbanicity, and Poverty
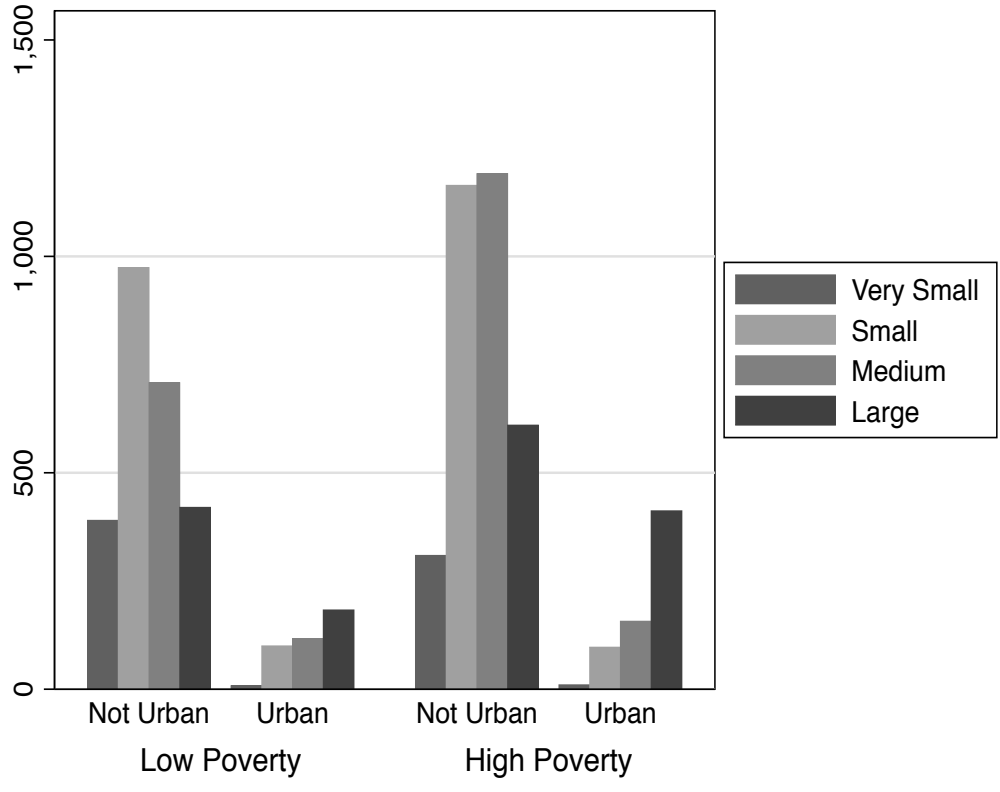
Figure 2: Boxplot of District-Specific ICCs by Grade, Grade-specific District Size, and Subject
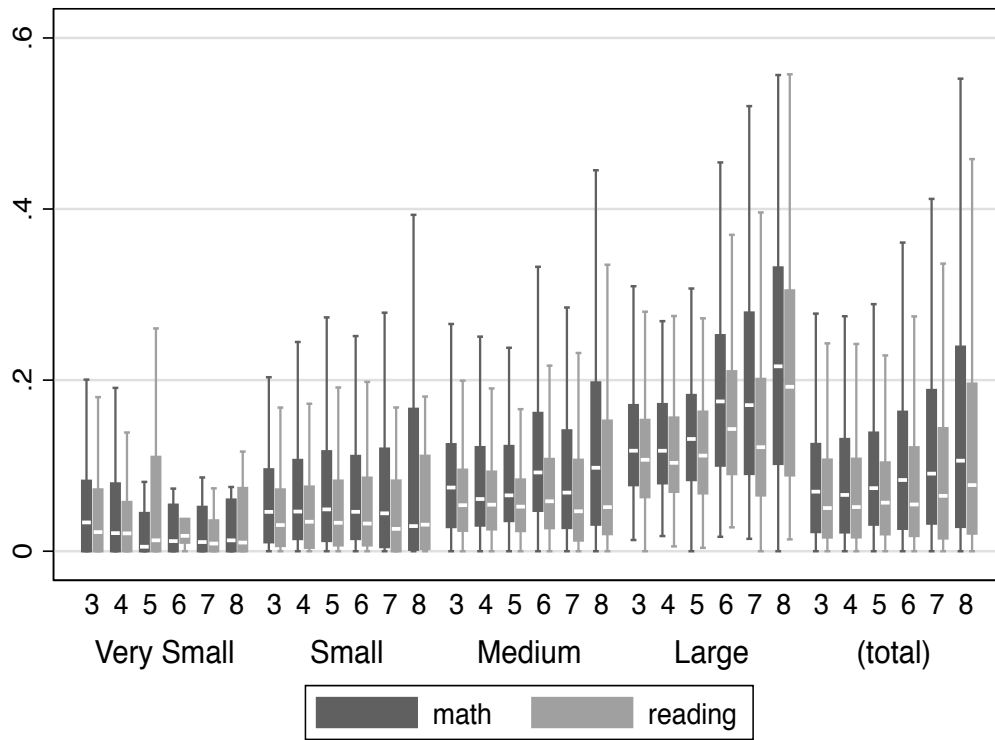
excludes outside values

Table 2: Results of random-effects meta analysis of within-district ICCs for mathematics achievement by district size and grade

Size	Schools	Grade	k	Mean ICC		τ	Empirical Percentiles		
							25th	50th	75th
All Districts		3	787	0.072	(0.004)	0.095*	0.022	0.070	0.125
		4	774	0.063	(0.003)	0.048*	0.023	0.067	0.131
		5	711	0.080	(0.004)	0.080*	0.032	0.075	0.140
		6	487	0.084	(0.005)	0.065*	0.028	0.086	0.170
		7	414	0.097	(0.005)	0.062*	0.033	0.094	0.190
		8	418	0.169	(0.015)	0.292*	0.032	0.119	0.267
Very	2-3	3	165	0.009	(0.004)	a	<0.001	0.034	0.084
Small	2-3	4	149	0.007	(0.003)	a	<0.001	0.022	0.079
Districts[b]	2	5	10	0.006	(0.019)	a	<0.001	0.025	0.081
	2	6	11	0.002	(0.006)	a	<0.001	0.014	0.073
	2	7	16	0.002	(0.005)	a	<0.001	0.016	0.110
	2	8	15	0.001	(0.005)	a	<0.001	0.018	0.075
Small	4-5	3	214	0.012	(0.003)	a	0.010	0.046	0.096
Districts[c]	4-5	4	216	0.030	(0.004)	0.030*	0.014	0.047	0.107
	3-5	5	313	0.014	(0.003)	a	0.013	0.049	0.117
	3-5	6	251	0.050	(0.006)	0.059*	0.014	0.047	0.113
	3	7	91	0.060	(0.010)	0.060*	0.005	0.048	0.137
	3	8	92	0.004	(0.002)	a	0.004	0.030	0.169
Medium	6-10	3	212	0.084	(0.012)	0.162*	0.029	0.075	0.125
Districts[d]	6-10	4	214	0.037	(0.003)	0.016	0.030	0.061	0.123
	6-11	5	214	0.040	(0.004)	0.018*	0.035	0.065	0.123
	6-11	6	133	0.058	(0.006)	0.027*	0.047	0.092	0.161
	4-6	7	150	0.027	(0.004)	0.014	0.027	0.072	0.141
	4-7	8	174	0.139	(0.014)	0.152*	0.032	0.099	0.210
Large	11+	3	172	0.118	(0.005)	0.055*	0.077	0.117	0.171
Districts[e]	11+	4	175	0.120	(0.005)	0.050*	0.079	0.117	0.172
	12+	5	155	0.141	(0.011)	0.122*	0.083	0.132	0.187
	12+	6	81	0.174	(0.011)	0.079*	0.100	0.175	0.252
	7+	7	126	0.183	(0.012)	0.107*	0.090	0.171	0.279
	8+	8	111	0.255	(0.043)	0.442*	0.107	0.235	0.352

Notes: a: τ estimated as 0, b: very small districts defined as the 10th percentile of size weighted by students served by grade, c: small districts defined as the 25th percentile of size weighted by students served by grade, d: medium districts defined as the 50th percentile of size weighted by students served by grade, e: large districts defined as the >50th percentile of size weighted by students served by grade, * $p(\tau = 0) < 0.05$, standard errors in parentheses.

Table 1: Number of Estimated ICCs and Sample Sizes

	Math						
	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	**Total**
Arkansas	33	33	21	16	12	11	**126**
	(15,668)	(15,251)	(11,307)	(8,756)	(7,904)	(7,252)	**(66,138)**
Arizona	73	73	71	54	44	44	**359**
	59,619	59,396	58,626	52,472	50,395	51,063	**331,571**
Colorado	49	49	47	37	32	33	**247**
	(47,340)	(46,570)	(45,448)	(41,666)	(40,890)	(40,402)	**(262,316)**
Florida	56	55	55	56	59	52	**333**
	(157,780)	(155,276)	(153,312)	(159,374)	(156,787)	(152,254)	**(934,783)**
Kansas	55	53	47	29	17	21	**222**
	(22,280)	(21,561)	(19,498)	(15,877)	(12,228)	(13,245)	**(104,689)**
Kentucky	89	89	88	45	28	27	**366**
	(35,960)	(36,418)	(36,163)	(25,814)	(21,558)	(21,317)	**(177,230)**
Louisiana	63	64	60	62	62	62	**373**
	(47,952)	(49,622)	(42,972)	(45,321)	(45,237)	(41,878)	**(272,982)**
Massachusetts	121	115	99	45	34	34	**448**
	(45,471)	(44,783)	(40,316)	(24,826)	(22,270)	(23,124)	**(200,790)**
North Carolina	96	94	92	77	71	72	**502**
	(96,468)	(93,134)	(91,233)	(85,836)	(83,089)	(83,455)	**(533,215)**
Wisconsin	102	98	83	31	20	20	**354**
	(34,944)	(34,460)	(30,880)	(19,134)	(16,426)	(16,670)	**(152,514)**
West Virginia	47	47	43	32	28	28	**225**
	(15,804)	(16,522)	(15,918)	(14,479)	(13,695)	(13,120)	**(89,538)**
Total	**784**	**770**	**706**	**484**	**407**	**404**	**3,555**
	(579,286)	**(572,993)**	**(545,673)**	**(493,555)**	**(470,479)**	**(463,780)**	**(3,125,766)**
	Reading						
Arkansas	33	32	21	16	12	11	**125**
	(15,633)	(14,964)	(11,285)	(8,744)	(7,885)	(7,239)	**(65,750)**
Arizona	73	73	71	54	44	44	**359**
	59,620	59,383	58,627	52,472	50,393	51,076	**331,571**
Colorado	49	49	47	37	32	33	**247**
	(46,179)	(46,451)	(45,412)	(41,638)	(40,854)	(40,381)	**(260,915)**
Florida	56	56	55	56	59	53	**335**
	(157,839)	(155,494)	(153,292)	(157,758)	(156,323)	(157,475)	**(938,181)**
Kansas	55	52	47	29	18	20	**221**
	(22,264)	(21,333)	(19,499)	(15,886)	(12,745)	(12,925)	**(104,652)**
Kentucky	89	90	88	45	28	27	**367**
	(35,960)	(36,523)	(36,163)	(25,814)	(21,558)	(21,317)	**(177,335)**
Louisiana	63	65	62	62	62	61	**375**
	(47,944)	(50,253)	(43,111)	(45,317)	(45,242)	(42,836)	**(274,703)**
Massachusetts	121	115	99	45	33	34	**447**
	(45,032)	(44,404)	(39,975)	(24,551)	(21,815)	(22,856)	**(198,633)**
North Carolina	96	94	92	77	72	72	**503**
	(96,158)	(92,848)	(90,986)	(85,592)	(83,290)	(83,225)	**(532,099)**
Wisconsin	102	98	83	31	20	19	**353**
	(34,785)	(34,372)	(30,795)	(19,081)	(16,368)	(16,148)	**(151,549)**
West Virginia	47	47	43	32	28	28	**225**
	(15,804)	(16,522)	(15,918)	(14,479)	(13,695)	(13,120)	**(89,538)**
Total	**784**	**771**	**708**	**484**	**408**	**402**	**3,557**
	(577,218)	**(572,547)**	**(545,063)**	**(491,332)**	**(470,168)**	**(468,598)**	**(3,124,926)**

Note: Number students in parentheses

Table 8: Results of random-effects meta analysis of low poverty within-district ICCs for mathematics achievement by district size and grade

Size	Schools	Grade	k	Mean ICC		τ	Empirical Percentiles		
							25th	50th	75th
All Districts		3	350	0.062	(0.007)	0.120*	0.017	0.056	0.102
		4	342	0.034	(0.003)	0.022*	0.014	0.045	0.094
		5	312	0.057	(0.004)	0.044*	0.018	0.063	0.122
		6	179	0.074	(0.007)	0.068*	0.018	0.066	0.138
		7	163	0.091	(0.007)	0.060*	0.024	0.086	0.184
		8	178	0.153	(0.015)	0.173*	0.029	0.093	0.259
Very	2-3	3	88	0.006	(0.004)	a	<0.001	0.027	0.075
Small	2-3	4	79	0.006	(0.004)	a	<0.001	0.011	0.047
Districts[b]	2	5	8	0.009	(0.024)	a	<0.001	0.044	0.222
	2	6	9	0.001	(0.006)	a	<0.001	0.014	0.073
	2	7	12	0.002	(0.006)	a	<0.001	0.012	0.069
	2	8	11	0.002	(0.006)	a	<0.001	0.007	0.075
Small	4-5	3	111	0.010	(0.003)	a	0.006	0.036	0.079
Districts[c]	4-5	4	109	0.012	(0.003)	a	0.014	0.036	0.081
	3-5	5	153	0.012	(0.003)	a	0.009	0.041	0.091
	3-5	6	97	0.044	(0.009)	0.065*	0.007	0.033	0.097
	3	7	34	0.081	(0.017)	0.074*	0.005	0.028	0.137
	3	8	38	0.014	(0.007)	0.016	0.008	0.034	0.133
Medium	6-10	3	81	0.078	(0.027)	0.233*	0.025	0.061	0.091
Districts[d]	6-10	4	88	0.022	(0.004)	a	0.021	0.045	0.080
	6-11	5	86	0.035	(0.005)	0.024*	0.018	0.058	0.114
	6-11	6	38	0.062	(0.012)	0.033	0.031	0.072	0.165
	4-6	7	52	0.006	(0.003)	a	0.027	0.056	0.103
	4-7	8	71	0.059	(0.009)	0.038*	0.027	0.064	0.149
Large	11+	3	62	0.104	(0.008)	0.044*	0.076	0.108	0.165
Districts[e]	11+	4	58	0.101	(0.008)	0.043*	0.060	0.109	0.173
	12+	5	56	0.136	(0.011)	0.068*	0.085	0.125	0.187
	12+	6	28	0.181	(0.021)	0.093*	0.093	0.135	0.335
	7+	7	52	0.205	(0.020)	0.119*	0.097	0.189	0.299
	8+	8	46	0.278	(0.033)	0.211*	0.150	0.277	0.369

Notes: a: τ estimated as 0, b: very small districts defined as the 10th percentile of size weighted by students served by grade, c: small districts defined as the 25th percentile of size weighted by students served by grade, d: medium districts defined as the 50th percentile of size weighted by students served by grade, e: large districts defined as the >50th percentile of size weighted by students served by grade, * p(τ = 0) < 0.05, standard errors in parentheses.

Table 7: Results of random-effects meta analysis of urban within-district ICCs for reading achievement by district size and grade

Size	Schools	Grade	k	Mean ICC		τ	Empirical Percentiles		
							25th	50th	75th
All Districts		3	104	0.105	(0.007)	0.057*	0.063	0.112	0.152
		4	106	0.117	(0.009)	0.071*	0.063	0.120	0.162
		5	102	0.105	(0.009)	0.070*	0.058	0.100	0.155
		6	90	0.081	(0.008)	0.049*	0.032	0.072	0.153
		7	84	0.083	(0.008)	0.048*	0.045	0.088	0.189
		8	84	0.121	(0.012)	0.084*	0.029	0.094	0.224
Very	2-3	3	--	--	--	--	.	.	.
Small	2-3	4	2	0.027	(0.163)	a	<0.001	0.015	0.030
Districts[b]	2	5	--	--	--	--	.	.	.
	2	6	--	--	--	--	.	.	.
	2	7	2	0.046	(0.103)	0.105	0.011	0.181	0.351
	2	8	2	0.016	(0.032)	a	0.013	0.157	0.302
Small	4-5	3	15	0.048	(0.017)	a	0.047	0.071	0.174
Districts[c]	4-5	4	15	0.132	(0.037)	0.116*	0.022	0.083	0.183
	3-5	5	13	0.009	(0.009)	0.010	0.011	0.031	0.140
	3-5	6	33	0.005	(0.003)	a	0.007	0.022	0.070
	3	7	12	0.004	(0.005)	a	0.023	0.058	0.090
	3	8	11	0.005	(0.006)	a	0.004	0.042	0.104
Medium	6-10	3	20	0.038	(0.010)	a	0.034	0.091	0.133
Districts[d]	6-10	4	17	0.059	(0.016)	0.037	0.035	0.071	0.125
	6-11	5	21	0.040	(0.010)	a	0.034	0.069	0.089
	6-11	6	21	0.059	(0.013)	a	0.046	0.061	0.122
	4-6	7	26	0.013	(0.006)	a	0.015	0.048	0.096
	4-7	8	32	0.090	(0.020)	0.087*	0.022	0.042	0.158
Large	11+	3	65	0.124	(0.009)	0.059*	0.088	0.122	0.164
Districts[e]	11+	4	66	0.126	(0.009)	0.060*	0.088	0.136	0.170
	12+	5	62	0.129	(0.010)	0.065*	0.087	0.131	0.163
	12+	6	31	0.153	(0.016)	0.062*	0.105	0.168	0.219
	7+	7	39	0.153	(0.019)	0.103*	0.084	0.132	0.239
	8+	8	35	0.182	(0.023)	0.119*	0.089	0.160	0.259

Notes: a: τ estimated as 0, b: very small districts defined as the 10th percentile of size weighted by students served by grade, c: small districts defined as the 25th percentile of size weighted by students served by grade, d: medium districts defined as the 50th percentile of size weighted by students served by grade, e: large districts defined as the >50th percentile of size weighted by students served by grade, * p(τ = 0) < 0.05, standard errors in parentheses.

Table 5: Results of random-effects meta analysis of non-urban within-district ICCs for reading achievement by district size and grade

| Size | Schools | Grade | k | Mean ICC | | τ | Empirical Percentiles | | |
							25th	50th	75th
All Districts		3	683	0.037	(0.002)	0.026*	0.012	0.042	0.093
		4	668	0.046	(0.003)	0.049*	0.012	0.047	0.091
		5	609	0.063	(0.009)	0.196*	0.016	0.052	0.094
		6	397	0.045	(0.003)	0.036*	0.016	0.050	0.114
		7	330	0.063	(0.005)	0.055*	0.013	0.060	0.134
		8	335	0.148	(0.017)	0.284*	0.020	0.080	0.244
Very	2-3	3	164	0.008	(0.003)	a	<0.001	0.024	0.074
Small	2-3	4	147	0.007	(0.004)	a	<0.001	0.021	0.059
Districts[b]	2	5	9	0.003	(0.015)	a	<0.001	0.013	0.110
	2	6	10	0.005	(0.010)	a	0.010	0.018	0.088
	2	7	15	0.003	(0.007)	a	<0.001	0.009	0.073
	2	8	13	0.001	(0.004)	a	<0.001	0.008	0.074
Small	4-5	3	199	0.010	(0.003)	a	0.004	0.029	0.068
Districts[c]	4-5	4	201	0.010	(0.003)	a	0.004	0.028	0.069
	3-5	5	300	0.012	(0.002)	a	0.007	0.034	0.081
	3-5	6	218	0.009	(0.002)	a	0.007	0.033	0.089
	3	7	78	0.065	(0.014)	0.089*	<0.001	0.027	0.097
	3	8	81	0.017	(0.006)	0.023*	0.002	0.032	0.123
Medium	6-10	3	192	0.031	(0.003)	0.018*	0.023	0.053	0.085
Districts[d]	6-10	4	197	0.057	(0.007)	0.079*	0.025	0.054	0.089
	6-11	5	193	0.018	(0.002)	a	0.023	0.051	0.082
	6-11	6	112	0.065	(0.009)	0.058*	0.025	0.058	0.106
	4-6	7	123	0.014	(0.003)	a	0.012	0.046	0.124
	4-7	8	142	0.077	(0.009)	0.065*	0.020	0.059	0.175
Large	11+	3	107	0.093	(0.006)	0.046*	0.061	0.096	0.148
Districts[e]	11+	4	109	0.097	(0.006)	0.050*	0.058	0.090	0.154
	12+	5	93	0.120	(0.036)	0.348*	0.061	0.102	0.163
	12+	6	50	0.122	(0.011)	0.051*	0.081	0.119	0.210
	7+	7	86	0.114	(0.010)	0.063*	0.062	0.112	0.195
	8+	8	76	0.250	(0.054)	0.462*	0.102	0.224	0.361

Notes: a: τ estimated as 0, b: very small districts defined as the 10th percentile of size weighted by students served by grade, c: small districts defined as the 25th percentile of size weighted by students served by grade, d: medium districts defined as the 50th percentile of size weighted by students served by grade, e: large districts defined as the >50th percentile of size weighted by students served by grade, * p(τ = 0) < 0.05, standard errors in parentheses.

Table 6: Results of random-effects meta analysis of urban within-district ICCs for mathematics achievement by district size and grade

Size	Schools	Grade	k	Mean ICC		τ	Empirical Percentiles		
							25th	50th	75th
All Districts		3	104	0.119	(0.009)	0.068*	0.082	0.116	0.164
		4	106	0.120	(0.008)	0.058*	0.076	0.124	0.169
		5	102	0.120	(0.008)	0.058*	0.083	0.125	0.174
		6	90	0.099	(0.009)	0.056*	0.048	0.100	0.175
		7	84	0.093	(0.008)	0.047*	0.051	0.106	0.227
		8	84	0.120	(0.010)	0.055*	0.040	0.130	0.235
Very	2-3	3	--	--	--	--	.	.	.
Small	2-3	4	2	0.049	(0.179)	a	0.030	0.040	0.051
Districts[b]	2	5	--	--	--	--	.	.	.
	2	6	--	--	--	--	.	.	.
	2	7	2	0.066	(0.084)	a	0.044	0.188	0.332
	2	8	2	0.023	(0.041)	a	0.018	0.134	0.250
Small	4-5	3	15	0.020	(0.012)	a	0.042	0.084	0.118
Districts[c]	4-5	4	15	0.134	(0.036)	0.103*	0.049	0.134	0.193
	3-5	5	13	0.023	(0.012)	a	0.025	0.036	0.169
	3-5	6	33	0.005	(0.004)	a	0.017	0.043	0.087
	3	7	12	0.002	(0.004)	a	0.005	0.052	0.121
	3	8	11	0.001	(0.003)	a	0.020	0.038	0.133
Medium	6-10	3	20	0.069	(0.016)	0.040*	0.054	0.106	0.167
Districts[d]	6-10	4	17	0.038	(0.011)	a	0.026	0.095	0.129
	6-11	5	21	0.069	(0.013)	a	0.062	0.086	0.123
	6-11	6	21	0.070	(0.014)	a	0.071	0.096	0.148
	4-6	7	27	0.004	(0.003)	a	0.019	0.061	0.112
	4-7	8	32	0.031	(0.009)	0.019*	0.027	0.088	0.155
Large	11+	3	65	0.138	(0.009)	0.052*	0.098	0.126	0.172
Districts[e]	11+	4	66	0.134	(0.008)	0.043*	0.092	0.133	0.173
	12+	5	62	0.143	(0.009)	0.050*	0.094	0.142	0.182
	12+	6	31	0.177	(0.014)	0.049*	0.107	0.192	0.244
	7+	7	39	0.172	(0.017)	0.085*	0.090	0.152	0.264
	8+	8	35	0.204	(0.021)	0.102*	0.095	0.211	0.296

Notes: a: τ estimated as 0, b: very small districts defined as the 10th percentile of size weighted by students served by grade, c: small districts defined as the 25th percentile of size weighted by students served by grade, d: medium districts defined as the 50th percentile of size weighted by students served by grade, e: large districts defined as the >50th percentile of size weighted by students served by grade, * p(τ = 0) < 0.05, standard errors in parentheses.

Table 4: Results of random-effects meta analysis of non-urban within-district ICCs for mathematics achievement by district size and grade

Size	Schools	Grade	k	Mean ICC		τ	Empirical Percentiles		
							25th	50th	75th
All Districts		3	683	0.064	(0.005)	0.096*	0.019	0.063	0.116
		4	668	0.051	(0.003)	0.039*	0.019	0.057	0.116
		5	609	0.071	(0.004)	0.079*	0.027	0.066	0.132
		6	397	0.081	(0.005)	0.068*	0.026	0.079	0.165
		7	330	0.101	(0.006)	0.073*	0.027	0.088	0.188
		8	334	0.175	(0.021)	0.358*	0.029	0.117	0.277
Very	2-3	3	164	0.009	(0.004)	a	<0.001	0.035	0.085
Small	2-3	4	147	0.007	(0.003)	a	<0.001	0.020	0.080
Districts[b]	2	5	9	0.008	(0.023)	a	<0.001	0.044	0.081
	2	6	10	0.001	(0.006)	a	<0.001	0.017	0.073
	2	7	14	0.001	(0.005)	a	<0.001	0.011	0.086
	2	8	13	0.001	(0.005)	a	<0.001	0.007	0.060
Small	4-5	3	199	0.012	(0.003)	a	0.010	0.041	0.092
Districts[c]	4-5	4	201	0.026	(0.004)	0.025*	0.013	0.044	0.101
	3-5	5	300	0.014	(0.003)	a	0.012	0.050	0.114
	3-5	6	218	0.055	(0.007)	0.068*	0.014	0.047	0.128
	3	7	79	0.075	(0.014)	0.079*	0.005	0.045	0.137
	3	8	81	0.007	(0.004)	a	0.002	0.029	0.180
Medium	6-10	3	192	0.083	(0.013)	0.171*	0.028	0.072	0.121
Districts[d]	6-10	4	197	0.038	(0.004)	0.017	0.030	0.058	0.116
	6-11	5	193	0.039	(0.004)	0.019*	0.034	0.064	0.123
	6-11	6	112	0.058	(0.007)	0.030*	0.041	0.089	0.164
	4-6	7	123	0.034	(0.005)	0.012	0.027	0.076	0.151
	4-7	8	142	0.157	(0.020)	0.211*	0.034	0.101	0.232
Large	11+	3	107	0.106	(0.007)	0.052*	0.065	0.105	0.169
Districts[e]	11+	4	109	0.112	(0.007)	0.050*	0.069	0.107	0.168
	12+	5	93	0.139	(0.016)	0.145*	0.068	0.122	0.187
	12+	6	50	0.172	(0.016)	0.088*	0.096	0.164	0.272
	7+	7	87	0.188	(0.015)	0.114*	0.090	0.171	0.285
	8+	8	76	0.277	(0.054)	0.459*	0.120	0.263	0.377

Notes: a: τ estimated as 0, b: very small districts defined as the 10th percentile of size weighted by students served by grade, c: small districts defined as the 25th percentile of size weighted by students served by grade, d: medium districts defined as the 50th percentile of size weighted by students served by grade, e: large districts defined as the >50th percentile of size weighted by students served by grade, * $p(\tau = 0) < 0.05$, standard errors in parentheses.

Table 3: Results of random-effects meta analysis of within-district ICCs for reading achievement by district size and grade

Size	Schools	Grade	k	Mean ICC		τ	Empirical Percentiles		
							25th	50th	75th
All Districts		3	787	0.049	(0.002)	0.037*	0.017	0.050	0.107
		4	774	0.057	(0.003)	0.055*	0.017	0.053	0.108
		5	711	0.069	(0.007)	0.178*	0.020	0.057	0.104
		6	487	0.052	(0.003)	0.038*	0.018	0.058	0.122
		7	414	0.067	(0.004)	0.052*	0.015	0.065	0.147
		8	419	0.145	(0.013)	0.252*	0.023	0.083	0.235
Very	2-3	3	165	0.008	(0.003)	a	<0.001	0.023	0.073
Small	2-3	4	149	0.008	(0.004)	a	<0.001	0.021	0.059
Districts[b]	2	5	10	0.004	(0.015)	a	<0.001	0.035	0.169
	2	6	11	0.005	(0.010)	a	0.010	0.022	0.088
	2	7	17	0.003	(0.006)	a	<0.001	0.011	0.073
	2	8	15	0.001	(0.004)	a	<0.001	0.013	0.116
Small	4-5	3	214	0.011	(0.003)	a	0.006	0.031	0.072
Districts[c]	4-5	4	216	0.011	(0.003)	a	0.004	0.035	0.076
	3-5	5	313	0.009	(0.002)	a	0.007	0.033	0.082
	3-5	6	251	0.008	(0.002)	a	0.007	0.033	0.086
	3	7	90	0.058	(0.011)	0.073*	<0.001	0.028	0.097
	3	8	92	0.013	(0.005)	0.018*	0.003	0.033	0.121
Medium	6-10	3	212	0.032	(0.003)	0.017*	0.024	0.054	0.095
Districts[d]	6-10	4	214	0.058	(0.007)	0.077*	0.026	0.054	0.094
	6-11	5	214	0.019	(0.002)	a	0.024	0.052	0.084
	6-11	6	133	0.065	(0.008)	0.055*	0.027	0.059	0.109
	4-6	7	149	0.014	(0.003)	a	0.013	0.048	0.107
	4-7	8	174	0.080	(0.008)	0.069*	0.020	0.056	0.169
Large	11+	3	172	0.105	(0.005)	0.052*	0.063	0.107	0.153
Districts[e]	11+	4	175	0.108	(0.005)	0.055*	0.069	0.103	0.156
	12+	5	155	0.125	(0.024)	0.298*	0.068	0.112	0.163
	12+	6	81	0.135	(0.009)	0.057*	0.090	0.143	0.210
	7+	7	125	0.127	(0.009)	0.071*	0.065	0.122	0.201
	8+	8	111	0.230	(0.042)	0.431*	0.094	0.195	0.337

Notes: a: τ estimated as 0, b: very small districts defined as the 10th percentile of size weighted by students served by grade, c: small districts defined as the 25th percentile of size weighted by students served by grade, d: medium districts defined as the 50th percentile of size weighted by students served by grade, e: large districts defined as the >50th percentile of size weighted by students served by grade, * $p(τ = 0) < 0.05$, standard errors in parentheses.

Table 11: Results of random-effects meta analysis of high-poverty within-district ICCs for reading achievement by district size and grade

Size	Schools	Grade	k	Mean ICC		τ	Empirical Percentiles		
							25th	50th	75th
All Districts		3	438	0.062	(0.003)	0.044*	0.024	0.062	0.124
		4	432	0.067	(0.004)	0.049*	0.022	0.064	0.122
		5	397	0.079	(0.013)	0.240*	0.023	0.064	0.120
		6	308	0.057	(0.004)	0.033*	0.025	0.061	0.128
		7	250	0.065	(0.005)	0.044*	0.016	0.072	0.163
		8	239	0.153	(0.024)	0.342*	0.026	0.102	0.245
Very	2-3	3	77	0.009	(0.006)	a	<0.001	0.028	0.086
Small	2-3	4	70	0.011	(0.007)	a	<0.001	0.027	0.081
Districts[b]	2	5	2	0.049	(0.115)	a	0.013	0.137	0.260
	2	6	2	0.003	(0.020)	a	0.001	0.019	0.038
	2	7	5	0.001	(0.010)	a	<0.001	<0.001	0.164
	2	8	4	0.002	(0.009)	a	0.004	0.041	0.095
Small	4-5	3	103	0.019	(0.005)	a	0.013	0.038	0.103
Districts[c]	4-5	4	107	0.025	(0.006)	0.026*	0.003	0.035	0.094
	3-5	5	159	0.015	(0.004)	a	0.009	0.047	0.103
	3-5	6	154	0.014	(0.003)	a	0.011	0.040	0.095
	3	7	56	0.014	(0.007)	a	0.002	0.028	0.099
	3	8	54	0.009	(0.006)	a	<0.001	0.042	0.124
Medium	6-10	3	132	0.030	(0.004)	a	0.030	0.055	0.104
Districts[d]	6-10	4	126	0.035	(0.004)	a	0.030	0.057	0.107
	6-11	5	127	0.017	(0.003)	a	0.019	0.049	0.086
	6-11	6	95	0.031	(0.005)	a	0.029	0.060	0.104
	4-6	7	97	0.020	(0.005)	0.012	0.013	0.052	0.108
	4-7	8	102	0.113	(0.014)	0.098*	0.024	0.072	0.196
Large	11+	3	110	0.113	(0.007)	0.056*	0.072	0.120	0.153
Districts[e]	11+	4	117	0.114	(0.007)	0.057*	0.074	0.116	0.156
	12+	5	99	0.130	(0.034)	0.338*	0.068	0.114	0.163
	12+	6	53	0.149	(0.012)	0.061*	0.095	0.165	0.206
	7+	7	72	0.132	(0.012)	0.077*	0.071	0.118	0.195
	8+	8	64	0.222	(0.060)	0.470*	0.099	0.176	0.338

Notes: a: τ estimated as 0, b: very small districts defined as the 10th percentile of size weighted by students served by grade, c: small districts defined as the 25th percentile of size weighted by students served by grade, d: medium districts defined as the 50th percentile of size weighted by students served by grade, e: large districts defined as the >50th percentile of size weighted by students served by grade, * $p(\tau = 0) < 0.05$, standard errors in parentheses.

Table 10: Results of random-effects meta analysis of high-poverty within-district ICCs for mathematics achievement by district size and grade

Size	Schools	Grade	k	Mean ICC		τ	Empirical Percentiles 25th	50th	75th
All Districts		3	437	0.076	(0.004)	0.054*	0.030	0.082	0.147
		4	432	0.088	(0.004)	0.062*	0.037	0.088	0.154
		5	399	0.094	(0.006)	0.102*	0.041	0.086	0.157
		6	308	0.089	(0.006)	0.054*	0.038	0.093	0.178
		7	251	0.104	(0.007)	0.067*	0.042	0.098	0.195
		8	240	0.179	(0.026)	0.382*	0.037	0.129	0.269
Very Small Districts[b]	2-3	3	77	0.016	(0.007)	a	0.004	0.041	0.139
	2-3	4	70	0.013	(0.007)	a	0.006	0.053	0.106
	2	5	2	0.000	(0.034)	a	<0.001	0.003	0.005
	2	6	2	0.013	(0.029)	a	0.010	0.032	0.054
	2	7	4	0.001	(0.010)	a	0.007	0.074	0.233
	2	8	4	0.001	(0.008)	a	0.016	0.047	0.246
Small Districts[c]	4-5	3	103	0.016	(0.005)	a	0.018	0.063	0.113
	4-5	4	107	0.061	(0.010)	0.063*	0.014	0.065	0.134
	3-5	5	160	0.018	(0.004)	a	0.023	0.066	0.155
	3-5	6	154	0.033	(0.005)	0.016	0.023	0.066	0.131
	3	7	57	0.020	(0.008)	0.020	0.008	0.053	0.130
	3	8	54	0.002	(0.003)	a	0.001	0.028	0.202
Medium Districts[d]	6-10	3	131	0.045	(0.005)	0.019	0.034	0.084	0.143
	6-10	4	126	0.055	(0.006)	0.027*	0.037	0.080	0.146
	6-11	5	128	0.040	(0.004)	a	0.042	0.070	0.138
	6-11	6	95	0.056	(0.007)	0.025	0.060	0.096	0.152
	4-6	7	98	0.052	(0.008)	0.027	0.027	0.086	0.175
	4-7	8	103	0.172	(0.028)	0.255*	0.039	0.123	0.233
Large Districts[e]	11+	3	110	0.125	(0.007)	0.058*	0.080	0.127	0.176
	11+	4	117	0.129	(0.006)	0.049*	0.086	0.129	0.170
	12+	5	99	0.141	(0.015)	0.140*	0.079	0.132	0.187
	12+	6	53	0.171	(0.013)	0.068*	0.109	0.183	0.244
	7+	7	74	0.170	(0.015)	0.101*	0.086	0.157	0.260
	8+	8	65	0.240	(0.060)	0.476*	0.095	0.211	0.325

Notes: a: τ estimated as 0, b: very small districts defined as the 10th percentile of size weighted by students served by grade, c: small districts defined as the 25th percentile of size weighted by students served by grade, d: medium districts defined as the 50th percentile of size weighted by students served by grade, e: large districts defined as the >50th percentile of size weighted by students served by grade, * p(τ = 0) < 0.05, standard errors in parentheses.

Table 9: Results of random-effects meta analysis of low poverty within-district ICCs for reading achievement by district size and grade

Size	Schools	Grade	k	Mean ICC		τ	Empirical Percentiles 25th	50th	75th
All Districts		3	349	0.031	(0.003)	0.024*	0.008	0.039	0.084
		4	342	0.046	(0.004)	0.057*	0.010	0.040	0.085
		5	314	0.041	(0.003)	0.027*	0.019	0.052	0.088
		6	179	0.045	(0.005)	0.040*	0.012	0.047	0.114
		7	164	0.067	(0.007)	0.056*	0.014	0.059	0.141
		8	180	0.130	(0.013)	0.152*	0.016	0.065	0.210
Very	2-3	3	88	0.007	(0.004)	a	<0.001	0.019	0.064
Small	2-3	4	79	0.006	(0.004)	a	<0.001	0.016	0.047
Districts[b]	2	5	8	0.003	(0.015)	a	<0.001	0.030	0.140
	2	6	9	0.006	(0.011)	a	0.012	0.022	0.088
	2	7	12	0.005	(0.009)	a	0.005	0.012	0.055
	2	8	11	0.001	(0.005)	a	<0.001	0.013	0.188
Small	4-5	3	111	0.008	(0.003)	a	<0.001	0.024	0.050
Districts[c]	4-5	4	109	0.009	(0.003)	a	0.007	0.032	0.063
	3-5	5	154	0.007	(0.003)	a	0.005	0.031	0.067
	3-5	6	97	0.008	(0.003)	0.007	0.005	0.019	0.064
	3	7	34	0.088	(0.019)	0.086*	<0.001	0.031	0.097
	3	8	38	0.023	(0.009)	0.031*	0.004	0.028	0.104
Medium	6-10	3	80	0.034	(0.006)	0.030*	0.019	0.043	0.081
Districts[d]	6-10	4	88	0.059	(0.013)	0.105*	0.022	0.046	0.086
	6-11	5	87	0.027	(0.004)	0.010	0.025	0.054	0.080
	6-11	6	38	0.098	(0.019)	0.096*	0.019	0.053	0.164
	4-6	7	52	0.012	(0.004)	a	0.012	0.040	0.091
	4-7	8	72	0.034	(0.007)	0.030*	0.014	0.040	0.145
Large	11+	3	62	0.086	(0.007)	0.038*	0.057	0.091	0.163
Districts[e]	11+	4	58	0.096	(0.008)	0.049*	0.046	0.087	0.156
	12+	5	56	0.105	(0.008)	0.042*	0.062	0.109	0.168
	12+	6	28	0.112	(0.013)	0.048*	0.073	0.098	0.215
	7+	7	53	0.120	(0.013)	0.066*	0.065	0.131	0.221
	8+	8	47	0.239	(0.029)	0.186*	0.089	0.212	0.334

Notes: a: τ estimated as 0, b: very small districts defined as the 10th percentile of size weighted by students served by grade, c: small districts defined as the 25th percentile of size weighted by students served by grade, d: medium districts defined as the 50th percentile of size weighted by students served by grade, e: large districts defined as the >50th percentile of size weighted by students served by grade, * p(τ = 0) < 0.05, standard errors in parentheses.