Abstract Title Page

Title: Intraclass Correlations for Three-Level Multi-Site Cluster-Randomized Trials of Science Achievement

Authors and Affiliations:
Carl D. Westine
Assistant Professor of Educational Research
University of West Georgia
cwestine@westga.edu

Background / Context:

In recent years, the impetus on experiments for educational research and evaluation has particularly revolved around experiments that involve clustering (Spybrook & Raudenbush, 2009; Institute of Education Sciences, 2013). A cluster-randomized trial (CRT) relies on random assignment of intact clusters to treatment conditions, such as classrooms or schools (Raudenbush & Bryk, 2002). One specific type of CRT, a multi-site CRT (MSCRT), is commonly employed in educational research and evaluation studies (Spybrook & Raudenbush, 2009; Spybrook, 2014; Bloom, Richburg-Hayes, & Black, 2007). The three-level MSCRT is a nested design with the level-three units (or sites) treated as a blocking variable, and the level-two units randomly assigned to treatment and control within each site.

As in all experimental studies, evaluators must design CRTs with appropriate power to detect an expected effect. A common challenge for evaluators planning CRTs is selecting an appropriate intraclass correlation (ICC), an estimate of the percentage of total variance that exists at the group level, to accurately power the study. For studies with more than one level of nesting, multiple ICCs must be estimated. For a three-level MSCRT with treatment at level-two, the evaluator must specify the within-site ICC, since the between-site variance is removed by blocking (Konstantopoulos, 2008).

Empirically estimating ICCs for use in mathematics, reading, and science are a common trend in the education literature (Bloom, Richburg-Hayes, & Black, 2007; Hedges & Hedberg, 2007; Brandon, Harrison, & Lawton, 2013; Hedges & Hedberg, 2013; Schochet, 2008; Jacob, Zhu, & Bloom, 2010; Westine, Spybrook, & Taylor, 2013). These estimates are typically based on completed evaluations, district datasets, and statewide databases. Until recently, the majority of these estimates were computed using two-level models (e.g., students nested in schools). However, many studies are being designed as MSCRTs with districts as sites and schools as the unit of randomization (Spybrook, 2014).

Purpose / Objective / Research Question / Focus of Study:

In this study, I aim to improve the design of MSCRTs for science achievement studies by producing estimates of within-district ICCs across all districts in an entire state. The result is a distribution of within-district ICCs which can be used to power a three-level MSCRT with treatment at the school level. Currently, evaluators planning trials focused on science outcomes must estimate ICCs based on empirical estimates from two-level or three-level models that do not block on district (Zhu, Jacob, Bloom, & Xu, 2012; Westine, Spybrook, & Taylor, 2013). The distribution of within-district ICCs serves as an empirical basis for the selection of an ICC value in order to facilitate better designs of MSCRTs in science education. Recent empirical work for mathematics and reading outcomes by Hedberg and Hedges (2011) suggests that distributions of within-district ICCs for states are asymmetrical. I examine if this holds for the outcome of science achievement.

Additionally, I investigate how an evaluator would utilize the distributional information to estimate a within-district ICC for a MSCRT design. In particular, an evaluator must select a point estimate to summarize the variances of participating districts. This estimate is needed in order to perform a power analysis, but such analyses typically occur before districts are even recruited. This analysis focuses on investigating whether within-district ICC estimates differ for (1) MSCRTs that include only a few districts with a larger number of schools per district; and (2) MSCRTs that include several more districts with a smaller number of schools per district. Using

actual student outcomes, I empirically investigate how the structure of an MSCRT impacts ICC estimates.

In summary, the following research questions guide this investigation:

1. What is the distribution of within-district ICCs for science education?
2. Does the number of districts in an MSCRT affect the mean within-district ICC?

Population / Participants / Subjects:

Data from the Texas Education Agency (TEA) for the academic year 2010-2011 is used for this study. The dataset includes student-level achievement data for science from the Texas Assessment of Knowledge and Skills (TAKS), student demographic information, and school and district identifiers. As in many other states, in Texas, science is tested in grades 5, 8, 10, and 11.

Significance / Novelty of study:

The specific choice of districts to include can significantly affect the number of schools per district needed to appropriately power a study because sample sizes are impacted by the within-district ICC. This gives rise to the notion of evaluators developing ways to improve MSCRT designs according to desired purposes (e.g., IES goal 3 or goal 4 studies) by strategically recruiting districts for their designs. However, there has been little empirical research with regard to specific strategies for district selection in three-level MSCRTs, and how this affects within-district ICCs because empirical examples of ICCs from large state databases, which enable examinations across sets of districts, are relatively recent, and none have looked at MSCRTs for science studies. This study contributes to this discussion and extends the discussion to the context of science education.

Statistical, Measurement, or Econometric Model:

The primary design examined is the three-level MSCRT with districts treated as sites and schools randomly assigned within sites. However, within-district ICCs are estimated using an unconditional two-level HLM for each individual district. In the interest of space, I present only the model for a two-level CRT.

To empirically estimate ICCs for each district I utilize a two-level HLM for each district. The unconditional model for the two-level HLM with students (Level 1) nested in schools (Level 2) is as follows. The Level 1 or student-level model is:

$$Y_{ij} = \beta_{0j} + r_{ij} \qquad r_{ij} \sim N(0, \sigma^2), \qquad \textbf{[1]}$$

where Y_{ij} is the outcome for individual $i \in \{1, \ldots, n_j\}$ in school $j \in \{1, \ldots, J\}$, β_{0j} is the average achievement at school j, and r_{ij} is a random student effect, which is assumed to be normally distributed with a mean of 0 and homogeneous variance σ^2. Therefore, σ^2 is the variance in achievement among students within schools. The Level 2 or school-level model is:

$$\beta_{0j} = \gamma_{00} + u_{0j} \qquad u_{0j} \sim N(0, \tau_{00}), \qquad \textbf{[2]}$$

where γ_{00} is the grand mean, and u_{oj} is a random school effect, which is assumed to be normally distributed with a mean of 0 and homogeneous variance τ_{00}. Therefore, τ_{00} is the variance in mean achievement among schools. A single ICC represents the proportion of total variance that exists among schools,

$$\rho = \frac{\tau_{00}}{\tau_{00} + \sigma^2}. \quad \textbf{[3]}$$

Data Collection and Analysis:

With the Texas dataset I first create a distribution of school-level ICCs for each district in the state with at least four schools and $n \geq 25$ students per school, using the two-level model. Using Stata (StataCorp, 2011), I execute LONEWAY by grade (for grades 5, 8 10, and 11) with school as a random factor to compute variance estimates for each district. The choice to use $J \geq 4$ is based on an initial investigation across all districts, and in response to findings from Hedberg and Hedges (2011) that districts with only a few schools can produce considerable variance in the school-level ICC. The result is a distribution of unconditional within-district ICCs for each grade.

Next, I investigate the variability of within-district ICCs across districts by generating and comparing confidence intervals on the mean within-district ICC for MSCRTs of different sizes. The two broader classes of MSCRTs, based on size, that are commonly used in the education literature are operationalized as follows. For a MSCRT with only a few districts I use $K = 3$ districts with a corresponding value of $J \geq 20$ schools per district. For a MSCRT with many districts, I use $K = 10$ with a corresponding value of $6 \leq J < 20$. The number of schools per district is used to identify sets of eligible districts for the different types of MSCRTs.

Considering the sets of eligible districts, I explore the range of within-district ICC values that could occur in a design for each grade in which science is tested. For each grade, I test if there is a difference in mean within-district ICC for districts with $J \geq 20$, $\bar{\rho}_{25,20,3}$, and districts with $6 \leq J < 20$, $\bar{\rho}_{25,6,10}$. Formally, this test is written a follows:

$$H_0: \bar{\rho}_{25,20,3} - \bar{\rho}_{25,6,10} = 0 \qquad H_A: \bar{\rho}_{25,20,3} - \bar{\rho}_{25,6,10} \neq 0\,.$$

Findings / Results:

In Texas, there are 154 districts with four or more schools that include fifth grade, 84 districts with four or more schools that include eighth grade, 50 districts with four or more schools that include tenth grade, and 51 districts with four or more schools that include eleventh grade. In Figure 1, I present the distribution of school-level ICCs for each district by grade using a bandwidth of 0.02. In order to plot all grades on the same graph, the percentage (rather than the count) of districts meeting the corresponding ICC level is shown.

(Insert Figure 1 about here.)

The distributions are fairly consistent across grades as well, as can be seen in Table 1. The mean within-district ICC for each grade ranges between 0.0781 and 0.0982. An F-test using analysis of variance under equal variances shows no significant difference (p=0.2610) in mean within-district ICC across grades.

(Insert Table 1 about here.)

Conceptually, the number of districts in an MSCRT does not change the underlying variance structure of the data. However, this choice does affect the number of districts eligible for a study, and therefore the sampling frame of districts for MSCRTs of various configurations

can be quite different. Respectively, in Grades 5, 8, 10, and 11, there are 46, 4, 2, and 2 districts meeting the sample size requirements for an MSCRT with $K = 3$ districts, $J \geq 20$ schools per district, and $n \geq 25$ students per school. Similarly, across grades, there are 68, 49, 19, and 21 districts that meet the sample size requirements for the MSCRT with $K = 10$ districts, $6 \leq J < 20$ schools per district, and $n \geq 25$ students per school.

In Table 2, I present a comparison, by grade, of the mean within-district ICC for a MSCRT with many districts, and a MSCRT with only a few districts. In Grade 5, a significant difference exists in the mean within-district ICC for the two designs (p=0.0020). More specifically, I find for the design with only a few districts, $\bar{\rho}_{25,20,3} = 0.1295$ ($SE = 0.0098$), and for the design with many districts, $\bar{\rho}_{25,6,10} = 0.0843$ ($SE = 0.0069$). In Grades 8, 10 and 11, the ability to test for significant differences in the mean within-district ICC for the two designs is limited by the number of eligible districts.

(Insert Table 2 about here.)

Conclusions:

The two common MSCRT design types drastically, but uniquely limit the eligibility of districts for each design by grade. The findings in grade 5 demonstrate that ICC estimates for MSCRTs can be refined further in some cases.

When estimating a within-district ICC value for a MSCRT power analysis, the evaluator should note the size of the districts in the sample from which the estimate is derived, and plan accordingly. Estimates for a design with only a few districts and a large number of schools per district were significantly larger than for a design with many districts and a smaller number of schools per district. This would suggest that prioritizing the recruitment of smaller districts into studies would tend to reduce participation requirements. However, while this logic is beneficial to researcher designing IES goal 3 or similar-type studies where generalization is not prioritized, it does not fit with the requirements of IES goal 4 or similar-type studies that do prioritize generalization. The results of this study are from one large, and representative (in terms of the grades in which testing occurs) state. Educational systems in other states may be structured quite differently, and further work in this area is needed to empirically estimate design parameters for planning MSCRTs. It will also be useful to explore whether significant differences exist between mean within-district ICC values in higher grades.

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

Hedberg, E. C., & Hedges, L. V. (2011). An investigation of the within- and between-variance structures of academic achievement in Massachusetts. *Society for Research on Educational Effectiveness*. Washington, DC.

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. doi:10.1177/0193841X14529126

Institute of Education Sciences. (2013, May 2). *Request for Applicantions: Statistical Research and Methodology in Education, CFDA Number: 84.305D*. Retrieved from Institute of Education Sciences Web site: http://ies.ed.gov/funding/pdf/2014_84305D.pdf

Jacob, R., Zhu, P., & Bloom, H. S. (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. (2008). The power of the test for treatment effects in three-level block randomized designs. *Journal of Research on Educational Effectiveness, 1*(4), 265-288. doi:10.1080/19345740802328216

Raudenbush, S. W., & Bryk, A. S. (2002). *Heirarchical Linerar Models: Applications and Data Analysis Methods* (2nd ed.). Thousand Oaks, CA: Sage Publications, Inc.

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

Spybrook, J. (2014). Detecting Intervention Effects Across Context: An Examination of the Precision of Cluster Randomized Trials. *Journal of Experimental Education, 82*(3), 334-357. doi:10.1080/00220973.2013.813364

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 Educational Sciences. *Educational Evaluation and Policy Analysis, 31*(3), 298-318. doi:10.3102/01623737093395244

StataCorp. (2011). Stata Statistical Software: Version 12. College Station, TX, USA: StataCorp LP.

Westine, C. D., Spybrook, J. & Taylor, J. T. (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. *Education Evaluation and Policy Analysis, 34*(1), 45-68. doi:10.3102/0162373711423786

Appendix B. Tables and Figures

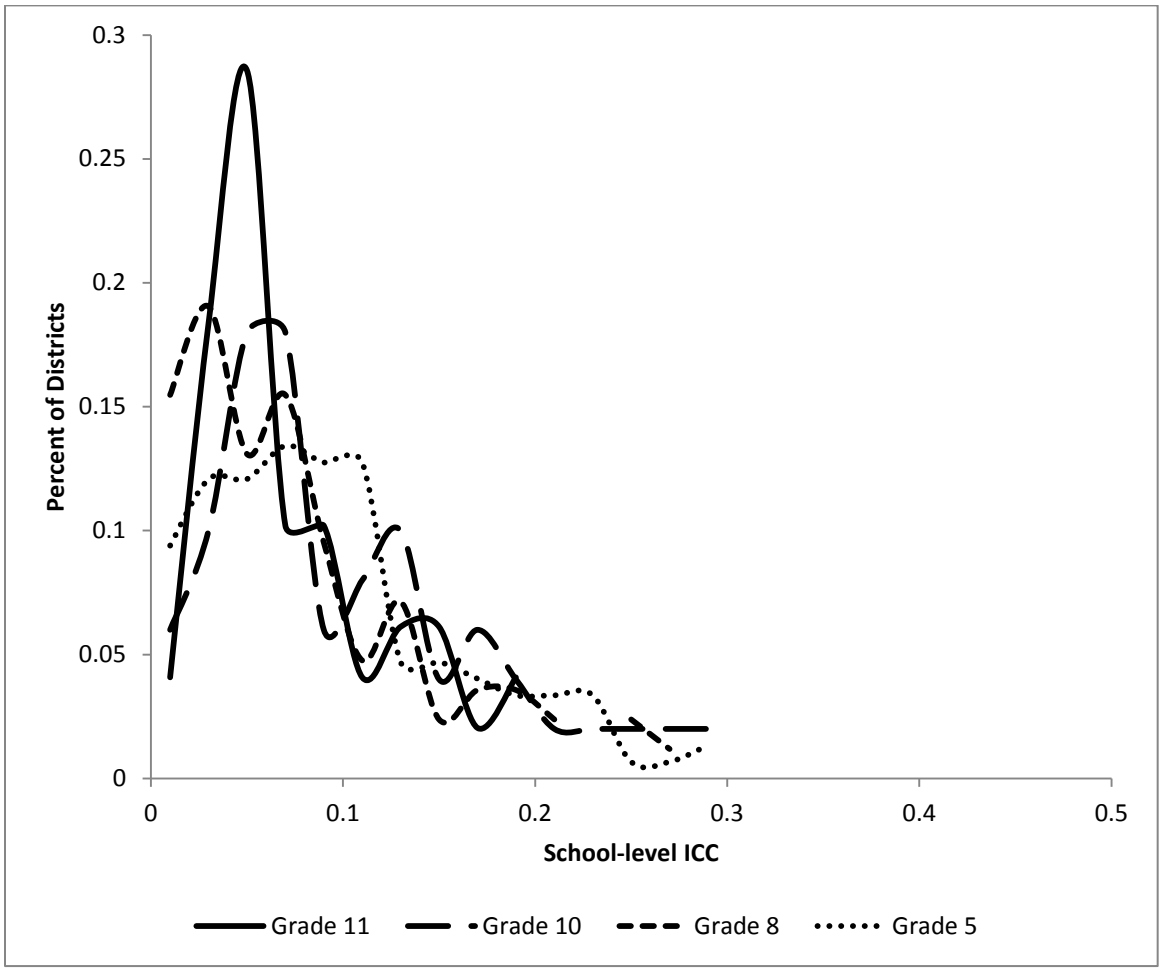

Figure 1. Distribution of unconditional school-level ICCs in science by grade for districts with four or more schools in Texas.

Table 1.

Average within-district ICC by grade for districts with $J \geq 4$

Grade	*K*	$\bar{\rho}$	*SE*
5	154	0.0964	0.0060
8	84	0.0781	0.0069
10	50	0.0982	0.0099
11	51	0.0933	0.0118
F(3,335)	1.340		
p	0.2601		

Table 2.

Comparison of mean ICC values by grade for MSCRTs with many districts and only a few districts

	MSCRT with many districts ($n \geq 25, 6 \leq J < 20$)			MSCRT with only a few districts ($n \geq 25, J \geq 20$)							
Grade	*K*	$\bar{\rho}_{25,6,10}$	*SE*	*K*	$\bar{\rho}_{25,20,3}$	*SE*	Difference	*SE*	d.f.	*t*	*p*
5	68	0.0843	0.0069	46	0.1295	0.0098	-0.0452	0.0116	112	3.9058	0.002
8	49	0.0877	0.0096	4	0.1102	0.0300	-0.0225	0.0347	51	0.6484	0.5196
10	19	0.0957	0.0153	2	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*
11	21	0.0893	0.0135	2	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*