4-2014

Improving the Design of Cluster-Randomized Trials in Education: Informing the Selection of Variance Design Parameter Values for Science Achievement Studies

Carl D. Westine
Western Michigan University, cwestine@hotmail.com

IMPROVING THE DESIGN OF CLUSTER-RANDOMIZED TRIALS IN EDUCATION: INFORMING THE SELECTION OF VARIANCE DESIGN PARAMETER VALUES FOR SCIENCE ACHIEVEMENT STUDIES

by

Carl D. Westine

A dissertation submitted to the Graduate College
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
Interdisciplinary Ph.D. in Evaluation
Western Michigan University
April 2014

Doctoral Committee:

Jessaca K. Spybrook, Ph.D., Chair
Chris L. S. Coryn, Ph.D.
Joseph A. Taylor, Ph.D.

IMPROVING THE DESIGN OF CLUSTER-RANDOMIZED TRIALS IN EDUCATION: INFORMING THE SELECTION OF VARIANCE DESIGN PARAMETER VALUES FOR SCIENCE ACHIEVEMENT STUDIES

Carl D. Westine, Ph.D.

Western Michigan University, 2014

The purpose of this three-essay dissertation is to provide practical guidance to evaluators planning cluster-randomized trials (CRTs) of science achievement. In an educational setting, interventions are often administered at the cluster level, while outcomes are typically measured at the student level through standardized achievement testing. When evaluating an intervention, a CRT is appropriate because it allows for treatment to be modeled at a different level than the unit of analysis, and properly accounts for the violation of independence that occurs due to nesting. Accurately designing a CRT involves estimating variance parameters (i.e., intraclass correlations [ICCs] and percent of variance explained [R^2] values). Prior efforts to improve the design of CRTs in education have primarily been limited to mathematics and reading disciplines, and the applicability of their findings to studies of science achievement is unknown.

I use three essays to present decision scenarios an evaluator faces when designing a CRT. In the first essay, the evaluator has limited information to inform the selection of ICCs for a three-level CRT. I use surface plots of relative efficiency to explore the robustness of an optimal design to misspecification of the ICCs. Findings suggest that

three-level CRTs are quite robust to misspecification of either or both ICCs. In the second essay, I resolve the challenge of limited information by using five years of achievement data from Texas to estimate ICCs for two- and three-level CRTs. I then analyze the decision of which covariate to include by estimating and evaluating R^2 values for demographic and pretest covariates. Findings suggest ICCs are larger in science than in mathematics and reading, and when a one-year lagged student-level science pretest is unavailable, a one-year lagged school-level science pretest is preferred. In the final essay, I recognize that a multi-site CRT (MSCRT) design is often more appropriate than a CRT, and the evaluator must once again select appropriate variance design parameter values. Using the Texas data, I empirically estimate a distribution of within-district ICCs, and show the number of districts in the MSCRT can impact the average within-district ICC value.

ACKNOWLEDGMENTS

There are numerous individuals that must be acknowledged for their contributions to this text and their personal support throughout graduate school and the dissertation process. Of course, I will start by thanking my committee chair, Jessaca Spybrook, and committee member Joseph Taylor for including me in their work and allowing me time to find ways to make my own contributions. I appreciate your leadership and professionalism in working together on this effort. In particular, I thank Jessaca for her continued dedication through the development and revision processes, which helped make my ideas come to fruition in a way that is both coherent and practical. I also thank Joe for helping to ensure my work is relevant.

I would like to extend a special thank you to my other committee member, Chris Coryn. Chris has been a constant source of inspiration throughout my training and dissertation work. He gave me my start, and I have learned so much from him about evaluation theory, practice, and methods—for which I am truly indebted. In addition to discussions about and revisions of my dissertation work, his influence as a mentor in my continued training is certainly relevant throughout the dissertation work.

There are several key people that contributed directly to the dissertation success, both methodologically and operationally. In particular, Eric Hedberg was instrumental in guiding me through the large task of working with and analyzing state datasets for my intended purposes. The analysis was certainly improved by his in-depth knowledge, for

which I am grateful and thankful. Additionally, several collaborators at BSCS have provided helpful input and important logistics in making the dissertation possible. Thank you to Karen Askinas, Steve Getty, Susan Kowalski, Molly Stuhlsatz, and Chris Wilson for helping to secure data and funding for the project, as well as providing valuable feedback on the execution of the studies. Finally, the data used in the dissertation was provided by the Texas Education Agency, and funding for the completion of this work was made possible by the National Science Foundation (grant #1118555). I thank each of these organizations for their respective contributions that helped make the work possible. I would also like to thank Hope Smith for editing and formatting the final draft.

Others played a less direct role in the dissertation, but certainly were essential ingredients in my successful completion of it. My program committee members, Magdalena Niewiadomska-Bugaj and Karen Vocke, also helped to shape my vision of evaluation and research. I thank them for helping to ensure my training was well-rounded and in the true spirit of an interdisciplinary program. I would like to extend an extra thank you to Karen for her encouragement as both a friend and mentor throughout my schooling.

I owe much thanks to both Arlen Gullickson and Lori Wingate for supporting several years of my graduate training through EvaluATE, which is also funded by the National Science Foundation (grants #0802245 and 1204683). In addition, thank you to the many individuals at the Evaluation Center, the Interdisciplinary Ph.D. in Evaluation program, and the Evaluation, Measurement, and Research program, which collectively

have provided ample opportunities to learn and develop skills as well as a place to come for answers. In particular, thank you, Mary Ramlow, for having one of just about everything, including answers to all the important questions about actually getting done.

As any Ph.D. student knows, the support of other students is essential. Many students in the Interdisciplinary Ph.D. in Evaluation program provided a much needed support system, which pushed me to successfully complete and defend the dissertation. While I do not want to leave any one individual out by naming names, I must thank Stephanie Evergreen, who personally recruited me to the program and showed me the ropes. Without her insistence, this journey would never have commenced. To everyone else, thank you for helping me learn and achieve my potential.

Lastly, I must express a humble thank you to all of my family and extended family for their encouragement and support over the years. Specific individuals deserve a special thank you, though thanks is hardly adequate. First, thank you to my dad, who instilled in me a love of learning and demonstrated by example how to earn the letters. Second, thank you to my mom, for her love and unconditional support of all my pursuits. Third, thank you to my kids, who sacrificed early in life to help me succeed. Finally, and most importantly, thank you to my wife, who followed me to this strange place called Kalamazoo, and showed incredible devotion to my personal achievement through her love, sacrifice, patience, and optimism over the years. Thank you all, I love you.

Carl D. Westine

TABLE OF CONTENTS

Table of Contents—Continued

CHAPTER

Table of Contents—Continued

CHAPTER

Table of Contents—Continued

LIST OF TABLES

List of Tables—Continued

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Background of the Problem

The value of a cluster-randomized trial (CRT) is hampered by a number of logistical and practical challenges stemming from the fact that the true power of a study is unknown until the conclusion of the experiment, but must be estimated at the beginning of the experiment for planning purposes. Evaluators rely on estimates of parameter values including effect sizes and variances (unconditional and conditional) to appropriately power designs. As individual disciplines like education increasingly move to test interventions using CRTs involving more complex hierarchical linear model (HLM) structures, the need for precise parameter value estimates through meta-analyses and empirical research to design high quality studies is heightened. Educational evaluators need practical guidance to ensure they can confidently and accurately specify design parameter values in their power analyses. This dissertation research contributes to this effort in the specific context of science education, where the demand for rigorous evaluations of achievement interventions is high, but the supply of empirical estimates of design parameters is low.

Currently, very few empirical examples of design parameters estimates exist, and many examples are needed to confidently enable the generalizability of design parameter estimates to new settings. For example, there is only one study with design parameter

estimates in science education (Zhu, Jacob, Bloom, & Xu, 2012), and its applicability is limited because it does not cover the range of grades or formats in which science is typically tested. Efforts to develop repositories of design parameter values (The University of Chicago Center for Advancing Research & Communication, 2011) have generated interest, but have largely been limited to mathematics and reading disciplines.

Over the past decade, significant advances in technology including the development of software have facilitated calculation of power for CRT designs, simplifying the process of appropriately powering these studies. Recent innovations in software development have focused on linking software to existing repositories of empirical inputs, for example, Optimal Design Plus (Spybrook, Bloom, Condon, Martinez, & Raudenbush, 2011). While these innovations will most certainly be useful, the reality is these projects have an infinite completion horizon. New research and evaluation questions are continually asked that require estimates of design parameter values that do not already exist. There is seemingly an endless need for additional empirical research, in conjunction with software enhancements, to facilitate better designs. In the absence of empirical research, the utility of software is diminished.

Due to a lack of empirical research on design parameter values, evaluators designing CRTs can face significant uncertainty in estimating these values for their studies. Three specific challenges associated with the design of CRTs are the focus of this dissertation:

- Challenge 1: Selecting parameter value estimates in the absence of precision.
- Challenge 2: Selecting the most effective covariate.

- Challenge 3: Selecting parameter value estimates for multi-site designs.

I elaborate on each of these challenges below. First, due to the high costs associated with conducting CRTs, achieving an optimal design (i.e., a design in which the variance of the treatment effect is minimized subject to a budget constraint) is highly desirable (Raudenbush, 1997). An overpowered design wastes valuable resources, while an underpowered design can render findings of limited usefulness. In the absence of reasonable parameter value estimates, it is important for evaluators and researchers to understand the implications of parameter value misspecification in order to better select values that maintain a balance between power and cost.

Second, meta-analytic and empirical estimates of design parameters for conducting CRTs help reduce the uncertainty associated with parameter value selection. Historically, the estimation of effect sizes through meta-analytic work has been the dominant approach leading to design parameter values of benefit for powering studies. As noted above, empirical estimates of variance parameters for multi-level studies rarely exist in the literature, and for many outcomes, empirical estimates do not exist at all. In the absence of variance estimates, evaluators and researchers often will borrow parameter value estimates from related disciplines where estimates are available, without regard to the applicability of these estimates to their context. For example, in science education where limited estimates exist, designers of CRTs for science achievement are often forced to borrow parameter value estimates to power studies from the mathematics and reading literature. Without design-specific parameter value estimates in science education, the likelihood for misspecification of parameter values is heightened.

Third, for efficiency purposes, many CRTs are designed with a site-level blocking variable and random assignment to treatment occurring within sites. The (within-site) variance design parameters required to power a multi-site CRT (MSCRT) are different than for a traditional CRT, and estimates must be produced in a slightly different way. Additionally, because the underlying true within-site variance in a MSCRT design is heavily influenced on the specific configuration of the relatively few sites recruited to the study, an evaluator's estimate of variance may be influenced by the number of sites in the study.

Experimental Designs in Education

In this section, I introduce the central underlying assumption for the research found in this dissertation. The assumption is that experimental designs are important in education and therefore worth improving. This motivates the need to design better CRTs in science education. I begin by describing the impetus on experimental research before describing the merits of CRTs. Detailed summaries of relevant literature for each specific essay are found in the individual chapters.

The Importance of Experimental Designs

The passing of the No Child Left Behind Act (NCLB) in 2001 and subsequent legislation including the Education Sciences Reform Act (ESRA) in 2002 marked a significant shift in federally funded educational research and evaluation organizations to one of evidence-based research (Institute of Education Sciences, 2013a). This priority continues today. For organizations such as the Institute of Education Sciences (IES),

which was established under the ESRA, priority was placed specifically on experimental studies that generated rigorous evidence about the effectiveness of educational programs, practices, and policies (Institute of Education Sciences, 2013b).

The experiment is the preferred method for establishing causal description (Shadish, Cook, & Campbell, 2002). However, some researchers have advocated that certain quasi-experiments (e.g., regression discontinuity designs, designs with carefully matched—focal local (Campbell, 1976)—comparison groups, and short interrupted-time series) are comparable to the experiment (Cook, Shadish, & Wong, 2008; Shadish, 2011). In certain cases, researchers have attempted to establish when quasi-experiments replicate the findings of experiments (Shadish, Clark, & Steiner, 2008). Yet, with few exceptions, experiments including CRTs remain the standard for causal research and large-scale evaluations at federal funding agencies like IES, the National Science Foundation (NSF), and the National Institutes of Health.

The Importance of CRTs in Education

In recent years, the impetus on experiments for educational research and evaluation has particularly revolved around experiments that involve clustering (Institute of Education Sciences, 2013b; Spybrook & Raudenbush, 2009). The applicability of CRTs for studying the effectiveness of educational programs is a result of the inherent nesting that occurs in the educational structure found in the United States (Raudenbush & Bryk, 2002). Students typically learn in traditional classroom environments, and these classrooms are located in schools, which are clustered in districts. Because educational material is most often delivered through the traditional classroom environment, treatment

is administered at the cluster level. Often individual schools or entire districts implement curriculum that is consistent across all classrooms within schools or schools within districts, thereby increasing the level at which the treatment is administered. Outcomes are typically measured at the student level through standardized achievement testing.

When evaluating the effects of these interventions, a CRT is appropriate because it allows for treatment to be modeled at a different level than the unit of analysis. The use of a CRT to test a social intervention will often produce more accurate effect size estimates than a traditional experiment (Hedges, 2007). Correctly modeling the nested structure can also increase the internal validity of the design by reducing the threat of contamination or treatment diffusion across treatment groups because the unit of randomization is, for example, an entire school as opposed to students or teachers within a school (Shadish et al., 2002). Furthermore, because treatment is administered collectively to groups rather than individuals, the standard assumption of independence that is necessary when statistically analyzing the experimental data in an ordinary least squares framework is violated. Using a hierarchical linear structure to model the data properly accounts for this violation of independence (Raudenbush & Bryk, 2002).

Powering Experiments and CRTs

The prevalence of CRTs in funded studies through the IES has been studied and shown to be increasing over time; however, many of the early funded studies were found to be inappropriately powered (Spybrook & Raudenbush, 2009). In terms of model structure, Hedges (2007) noted that researchers often fail to account for group effects when powering studies and, consequently, they overstate the precision of results.

Likewise, when three-levels are used, it is important to account for the nesting that occurs at the second level, or power overestimation will occur (Konstantopoulos, 2008; Moerbeek, 2004).

Experiments must be designed to be feasible within the constraint of budget, but they should also be designed optimally to maximize the power to detect a minimum detectible effect size (MDES), or likewise, to minimize the standard error of the treatment effect estimate (Raudenbush, 1997). The MDES is the smallest true effect a design can detect (Bloom, Richburg-Hayes, & Black, 2007). For a traditional experiment with a desired MDES and error tolerance, once variance across individuals has been estimated, the power calculation is driven by a single decision variable—the number of individuals. Since power is a monotonically increasing function with respect to individuals, power is maximized by using as many individuals as can be afforded.

When evaluators use a HLM structure, they face additional challenges choosing an appropriate design, including the choice of additional design parameters and maintaining optimality with respect to cost. Like a traditional experiment, CRTs need to be designed with sufficient power, such that the researcher is able to detect a statistically significant effect when one actually exists.

Conducting a two-level CRT requires a bit more sophistication in the planning stages than that of a traditional experiment because under the same set of assumptions (MDES, error tolerance, and variance estimation), there are two decision variables that drive the calculation of power—the number of individuals and the number of clusters. Consequently, the evaluator must specify the variance at each level of nesting. Variance

in CRTs is described using an intraclass correlation (ICC), which for a two-level design represents the percentage of total variance that exists at Level 2. In a two-level CRT, the specification of the Level 2 ICC necessarily determines the Level 1 or residual variance partition. As models add more levels of nesting, variance must be partitioned across each level for planning purposes. Unlike traditional experiments where power to detect a MDES under the estimation of variance is driven by the number of individuals, power for CRTs is mostly driven by the number of clusters in the design, and to a much lesser extent by the number of individuals in each cluster (Raudenbush & Liu, 2000). The logic can be extended to CRTs with additional levels of nesting as well, with power being driven by the highest cluster level (Konstantopoulos, 2008).

Improving the Design of CRTs

Efforts to improve the design of CRTs have focused on understanding the impacts of model structure (like those studies mentioned above), meta-analyses, estimating variance design parameters, and understanding the precision of these parameter values. In this dissertation, I focus on only the empirical estimation of variance design parameters (i.e., ICCs and R^2 values), noting that effect sizes, although also a design parameter, are generally estimated through meta-analytic approaches. Specific attention is placed not only on the empirical estimates, but also on the proper use of these estimates.

Dissertation Format and Related Purposes of the Three Studies

This dissertation consists of three essays, which together seek to improve the design of CRTs in education. The opening chapter orients the reader to the broader context and more specifically to the challenges regarding the selection of variance design parameter values that evaluators face when designing CRTs. The three essays appearing in this dissertation as Chapters II, III, and IV are briefly described below. A final chapter considers implications and limitations of this collection of research in an effort to highlight new directions for future research in this area.

Each essay is written to offer practical advice regarding particular challenges an evaluator faces when designing a three-level CRT. In the first essay, the evaluator must estimate ICCs for a traditional CRT, but is faced with limited information regarding the ICCs. In the second essay, estimates of ICCs are presented in an effort to resolve the challenge of limited information, but the evaluator must then decide which covariate to use. In the final essay, it is noted that in some situations a MSCRT is a more appropriate design, and the evaluator must select an appropriate within-site ICC from a distribution of values. Below, each essay is described in greater detail, and specific research questions for each essay are presented.

Essay 1

Overview. The first essay includes an efficiency analysis using a three-level HLM framework in order to understand the robustness of an optimal design for a three-level CRT to misspecifications of ICCs. Misspecification is undesirable because it

signifies a waste of precious resources. Unfortunately, when designing a CRT, misspecification of ICCs is common because the true ICC is not known until after the experiment has been conducted. Additionally, very few empirical estimates of ICCs in which to inform the evaluator's decision exist in the published literature. For some disciplines, like science education, empirical estimates do not exist, meaning misspecification, in theory, could be very large.

In this essay, a model for examining misspecification in a three-level context is derived using the foundations of optimal design for a three-level CRT (Konstantopoulos, 2009) as well as an efficiency analysis, using relative efficiency (RE), of optimal designs in a two-level CRT framework (Korendijk, Moerbeek, & Maas, 2010). An efficiency analysis for a three-level design is different from a two-level design in that there are two ICCs and three cost factors to consider. Each ICC can be either over-specified or under-specified, creating several relevant scenarios that must be considered. The underlying cost structure of adding participants at each level can also impact the efficiency of a design and the impact of misspecification. It is important for evaluators to understand how robust optimal designs are to misspecification of ICCs in order to minimize the cost associated with over-powering or under-powering a study.

Research objectives. The following research questions are addressed in the first essay:

1. What are the ranges of misspecification for Level 2 and Level 3 ICCs that maintain a high level of RE?

2. What are the implications on the range of acceptable misspecification for different costs of adding additional units?
3. What are the implications on the range of acceptable misspecification for different combinations of over-specification and under-specification between the ICC values?

Essay 2

Overview. In the second essay, a state database of achievement data from Texas is used to empirically estimate ICCs and R^2 values using a variety of pretest and demographic covariates for two-level and three-level CRTs in science education. This reduces the likelihood of design parameter misspecification in this context.

Recent research on empirically estimating design parameter values in education has focused on mathematics and reading outcomes, leaving evaluators of science achievement interventions, for example, whole school curricula, to borrow ICC and R^2 values from these other subjects without regard for the applicability of these estimates. Furthermore, because science is tested infrequently, only certain covariate options are available. Often the most recent student-level pretest covariate, which typically is the best predictor of student performance, is lagged two, three, or more years, and the most desirable alternative is not immediately obvious.

The results of the empirical estimation procedure for science are used to compare the applicability of ICC estimates from mathematics and reading. In addition, various covariate models are considered to explore the desirability of pretest and demographic covariates in the years in which science is tested.

Research objectives. The following research questions are addressed in the second essay:

1. What are unconditional ICCs for science achievement outcomes?
2. How do the empirical estimates of ICCs for science achievement compare to those for reading and mathematics achievement?
3. For the grade levels in which science is tested, which covariate sets explain the most variance?

Essay 3

Overview. The third essay expands on the second essay to develop empirical estimates of design parameters for three-level MSCRT designs of science achievement using the Texas data. In the MSCRT design considered, districts are treated as a blocking variable, and schools are randomly assigned to treatment and control within districts. Experiments of this form will often be utilized in educational evaluations because they are typically cheaper to conduct than a traditional CRT.

Because randomization occurs in schools within-districts, the ICC design parameter required for a MSCRT power analysis is different than for a traditional CRT. An accurate estimate of the within-district ICC is needed to appropriately power a MSCRT design. The within-district ICC is different from the school-level ICC in that district variance is explained through blocking.

As is true for CRTs in science education, the evaluator has limited access to empirical estimates of a relevant ICC value for a MSCRT design. Using a two-level model within each district in the state, a distribution of within-district ICCs is empirically

estimated. One method of estimating the ICC from the distribution is to take the average within-district ICC. However, the number of districts needed for an MSCRT design can vary depending on the purpose of the study. The true within-district ICC for the recruited districts of one study may differ significantly from one study to the next.

There are typically two types of MSCRTs used in practice: those with only a few districts, but a large number of schools per district, and those with many districts, but a small number of schools per district. By categorizing the districts by size, and therefore design, appropriate ICCs for each MSCRT design are estimated and compared.

Research objectives. The following research questions are addressed in the third essay:

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

Significance of the Research

The collection of essays presented in this dissertation push the boundary of empirical research on improving the design of cluster-randomized trials in education. In each essay, important questions that science education evaluators currently face are considered. Collectively, the three essays provide practical guidance to evaluators planning CRTs in education.

Several noteworthy contributions inform the selection of variance design parameter values for studies of science achievement. First, evaluation efficiency is a

topic that is touched on in each of the three essays. The dissertation serves as an important example for evaluation practitioners of how evaluative decisions like the specification of ICC values can impact the cost-effectiveness of an evaluation. Second, there is a focus on creating an empirical base of design parameters for the evaluation of science education interventions. Empirical estimates of design parameters for science education do not currently exist across the range of grades in which science is tested, and the results of this dissertation fill this void for traditional three-level CRTs as well as MSCRTs. Third, in the context of science, where annual testing is not the norm, the comparison of lagged pretests is relevant and important, and likely will serve as an example for other researchers as new subject areas are explored. Fourth, there are important distinctions between CRTs and MSCRTs, and evaluators must pay attention to subtle differences in designs when selecting variance design parameter values.

Other contributions are timely. For example, the notion of improved outcomes in science, technology, engineering, and mathematics (STEM) disciplines continues to be relevant to educational policy makers. The tackling of methodological research questions pertinent to STEM education evaluation helps to ensure science education evaluations are of the highest quality. Additionally, as educational evaluations increasingly involve more than two levels of nesting, the need for design parameters value estimates from three-level CRT and MSCRT models is especially relevant.

In the following three chapters, these and other contributions are described. In the closing chapter, I suggest ideas as to where the research can go next.

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

Campbell, D. T. (1976). Focal local indicators for social program evaluation. *Social Indicators Research, 3*, 237-256. doi:10.1007/BF00286305

Cook, T. D., Shadish, W. R., & Wong, V. C. (2008). Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons. *Journal of Policy Analysis and Management, 27*(4), 724-750. doi:10.1002/pam.20375

Hedges, L. V. (2007). Correcting a significance test for clustering. *Journal of Educational and Behavioral Statistics, 32*(2), 151-179. doi:10.3102/1076998606298040

Institute of Education Sciences. (2013a, May 14). *IES About US*. Retrieved from Institute of Education Sciences website: http://ies.ed.gov/aboutus/

Institute of Education Sciences. (2013b, May 2). *Request for applicantions: Statistical research and methodology in education, CFDA Number: 84.305D.* Retrieved from Institute of Education Sciences website: http://ies.ed.gov/funding/pdf/2014_84305D.pdf

Konstantopoulos, S. (2008). The power of the test for treatment effects in three level cluster randomized designs. *Journal of Research on Educational Effectiveness, 1*(2), 66-88. doi:10.1080/19345740701692522

Konstantopoulos, S. (2009). Incorporating cost in power analysis for three-level cluster-randomized designs. *Evaluation Review, 33*(4), 335-357. doi:10.1177/0193841X09337991

Korendijk, E. J., Moerbeek, M., & Maas, C. J. (2010). The robustness of designs for trials with nested data against incorrect initial intracluster correlation coefficient estimates. *Journal of Educational and Behavioral Statistics, 35*(5), 566-585. doi:10.3102/1076998609360774

Moerbeek, M. (2004). The consequence of ignoring a level of nesting in multilevel analysis. *Multivariate Behavioral Research, 39*(1), 129-149. doi:10.1207/s15327906mbr3901_5

Raudenbush, S. W. (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods, 2*(2), 173-185. doi:10.1037/1082-989X.2.2.173

Raudenbush, S. W., & Bryk, A. S. (2002). *Heirarchical linear Models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Raudenbush, S. W., & Liu, X. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological Methods, 5*(2), 199-213. doi:10.1037//1082-989X.5.2.199

Shadish, W. R. (2011). Randomized controlled studies and alternative designs in outcome studies : Challenges and opportunities. *Research on Social Work Practice, 21*(6), 636-643. doi:10.1177/1049731511403324

Shadish, W. R., Clark, M. H., & Steiner, P. M. (2008). Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random and nonrandom assignments. *Journal of the American Statistical Association, 103*(484), 1334-1356. doi:10.1198/016214508000000733

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston, MA: Houghton Mifflin.

Spybrook, J., Bloom, H., Condon, R., Martinez, A., & Raudenbush, S. W. (2011). *Optimal Design Plus empirical evidence: Documentation for the "Optimal Design" software.* Ann Arbor, MI: University of Michigan.

Spybrook, J. K., & 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. doi:10.3102/01623737093395244

The University of Chicago Center for Advancing Research & Communication. (2011). *Variance almanac (VA) of academic achievement*. Retrieved January 30, 2014, from Center for Advancing Research & Communication website: https://arcdata.uchicago.edu/

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

CHAPTER II

THE ROBUSTNESS OF OPTIMAL DESIGNS TO MISSPECIFICATION OF INTRACLASS CORRELATIONS FOR THREE-LEVEL CLUSTER-RANDOMIZED TRIALS IN EDUCATION

As the evidence-based movement pushes evaluators to increasingly utilize experimental designs including the cluster-randomized trial (CRT), evaluators must rely on a limited supply of empirically estimated design parameters to properly power their studies. For many educational outcomes, accurate design parameters are unavailable. In some situations design parameters are known for a particular population or a related outcome, and evaluators will borrow these estimates without regard to their applicability from one study to the next. When a design parameter is not accurately estimated, inefficiency in the form of over-powering or under-powering the study will occur, which can lead to a significant waste of resources. In disciplines like education where resources are highly scarce and the number of stakeholders is high, inefficiency in evaluations can be particularly problematic. It is important for evaluators to understand how the selection of particular design parameter values impacts the overall quality of the CRT design in order to maximize the efficiency of their designs.

When designing a CRT, one challenge for the evaluator is to correctly estimate design parameter values in order to maximize the chance of detecting an effect when one in fact exists. In this study, one particular design parameter is emphasized, the intraclass correlation (ICC), which measures the percent of total variance found at each level of nesting. Unfortunately, the true variance decomposition is not known until after the

experiment occurs and the data has been collected. Thus, there is a tremendous likelihood that misspecification of the true ICC value(s) will occur, and produce an inefficient design.

The purpose of this essay is to explore the robustness of optimal designs to misspecification of one or both ICC values in a three-level CRT. Korendijk, Moerbeek, and Maas (2010) first considered the impacts of misspecification of the ICC for CRT designs using a two-level model. However, the conclusions drawn from the two-level model, do not explicate the decision-making process for the evaluator designing a three-level study since there are two ICC values that must be specified under this model structure.

Korendijk et al. (2010) studied the impacts of ICC misspecification using relative efficiency (RE) as their metric. RE, which is discussed more formally below, is a comparison of how much larger the variance in the treatment effect estimate is for a model with an incorrectly specified ICC to that of its minimum possible value. Moreover, the reciprocal of RE measures the extent to which the sample size for a model with an incorrectly specified ICC would need to be increased in order to produce the same level variance of the treatment effect estimate had the estimated ICC been correct. Larger values of RE are preferred, meaning there is only a small difference in the variance of the treatment effect estimate between the two model specifications.

Findings from Korendijk et al. (2010) show that a high level of RE is maintained for a wider range of over-specification of values than for under-specification of values. As a rough estimate, $RE \geq 0.90$ is maintained when an initial ICC estimate falls within a

range of 25% (i.e., 75% under-specification) to 275% (i.e., 175% over-specification) of its true value. This indicates that the ICC is quite robust to misspecification in the two-level model. With only two levels in the model, the authors instruct, "Assuming that a researcher in pursuit of a reasonable estimate for the intracluster correlation coefficient value has obtained a range of plausible values, the conclusion can be drawn that it is best to choose a high value within the obtained range" (p. 575). However, while this advice is useful in cases when a range of plausible values is small, if ICCs are unknown or a range is sufficiently imprecise, significant over-specification may result in large amounts of unnecessary participation which ultimately wastes important resources.

In education, it is becoming more common for experimental designs to have three or even four levels of nesting (Spybrook, in press; Spybrook & Raudenbush, 2009). For example a three-level model could include students nested in teachers nested in schools, or students nested in schools nested in districts. Issues of efficiency must be investigated in these contexts as well because misspecification can occur for multiple ICCs. Since the power of a study is most influenced by the number of participants at the highest level of nesting, of particular interest is whether the range of acceptable misspecification for the ICC at highest level of nesting is different from the range of acceptable misspecification for the ICC in a two-level model. Additionally, in models with more than two levels the implications of misspecification of one ICC must be studied in the context of misspecification of other ICCs.

Empirical Estimates of ICCs in Three-Level Models

The focus of this study is on designs with three-levels with a particular emphasis on education. Several examples of Level 2 and Level 3 ICCs from three-level designs in education are found in Table 2.1; these ICCs are summarized to highlight the considerable amount of variability that can exist from one study to the next. Two important conclusions can be drawn from Table 2.1, and motivate the research questions in this study. First, ICC values typically fall on the range from 0.05 to 0.30. Second, there is no consistency between whether the Level 2 ICC is larger than the Level 3 ICC, or the opposite is true. These two conclusions can be attributed to the nature of the levels, outcome measure, and subject area.

Using empirical estimates from the literature as a guide, in this study, the following questions are asked of CRTs that utilize three-level nested models:

1. What are the ranges of misspecification for Level 2 and Level 3 ICCs that maintain a high level of RE?
2. What are the implications on the range of acceptable misspecification for different costs of adding additional units?
3. What are the implications on the range of acceptable misspecification for different combinations of over-specification and under-specification between the ICC values?

Table 2.1

Examples of Level 2 and Level 3 ICCs from Three-Level Models in Education

			Range of ICCs	
Nesting Structure	Subject(s)	Source	Level 2	Level 3
Students in Classrooms in Schools	Math/Reading	(Konstantopoulos, 2009)	0.06–0.14	0.10–0.28
Students in Classrooms in Schools	Math/Reading/Science	(Zhu, Jacob, Bloom, & Xu, 2012)	0.03–0.38	0.04–0.17
Students in Schools in Districts	Math/Reading	(Hedberg & Hedges, 2011)	0.09–0.11	0.07–0.11
Students in Schools in Districts	Math/Reading	(Hedges & Hedberg, in press)	0.055–0.418	0.001–0.132
Students in Schools in Districts	Science	(Westine, Spybrook, & Taylor, in press)	0.10–0.14	0.06–0.08

Methodology

In this section, I describe the methods used to address the research questions. I begin by presenting the three-level model for a CRT. This is followed by a discussion of how an optimal design is derived for a three-level model. Next, I formally define the measure of efficiency used to judge the robustness of designs in this study, RE (Korendijk et al., 2010). Finally, I outline various relationships between the ICCs that are important to consider in assessing RE in a three-level model.

Model

Presented below is the theoretical framework for use in a three-level CRT. To model a three-level CRT, I use a three-level hierarchical linear model (HLM). In the model, each Level 2 and Level 3 variable is treated as a random effect. For convenience, I refer to Level 1 units as students, Level 2 units as teachers, and Level 3 units as schools. However, the reader should note that the analysis is not limited to this particular nested structure.

The unconditional model for the three-level HLM with Level 1 students nested within Level 2 teachers nested within Level 3 schools is as follows. The Level 1 model is:

$$Y_{ijk} = \pi_{0jk} + e_{ijk} \qquad e_{ijk} \sim N(0, \sigma^2) \qquad \textbf{[1]}$$

where Y_{ijk} is the outcome for Level 1 student $i \in \{1, \ldots, n_{jk}\}$, in Level 2 teacher $j \in \{1, \ldots, J_k\}$, in Level 3 school $k \in \{1, \ldots, K\}$; π_{0jk} is the mean outcome of Level 2 teacher *j* in Level 3 school *k*; and e_{ijk} is a random Level 1 effect which is assumed to be normally distributed with mean 0 and homogenous variance σ^2. Therefore, σ^2 is the variance in outcome among Level 1 students within Level 2 teachers. The Level 2 model is:

$$\pi_{0jk} = \beta_{00k} + r_{0jk} \qquad r_{0jk} \sim N(0, \tau_\pi), \qquad \textbf{[2]}$$

where β_{00k} is the mean outcome of Level 3 school *k*, and r_{0jk} is the random Level 2 effect which is assumed to be normally distributed with mean 0 and homogenous variance τ_π. Therefore, τ_π is the variance in the mean outcome among Level 2 teachers within Level 3 schools. The Level 3 model is:

$$\beta_{00k} = \gamma_{000} + \gamma_{001}W_k + u_{00k} \qquad u_{00k} \sim N(0, \tau_\beta), \quad \textbf{[3]}$$

where γ_{000} is the grand mean, γ_{001} is the main effect of the treatment, W_k is the treatment contrast indicator that equals 0.5 for treatment and -0.5 for control, and u_{00k} is a random Level 3 effect which is assumed to be normally distributed with mean 0 and homogenous variance τ_β. Therefore τ_β is the variance in mean outcome among Level 3 schools.

In the three-level HLM, there are two ICCs. The Level 2 ICC, or proportion of total variance that exists among Level 2 teachers within Level 3 schools is

$$\rho_2 = \frac{\tau_\pi}{\tau_\beta + \tau_\pi + \sigma^2}. \quad \textbf{[4]}$$

The Level 3 ICC, or proportion of total variance that exists among Level 3 schools is

$$\rho_3 = \frac{\tau_\beta}{\tau_\beta + \tau_\pi + \sigma^2}. \quad \textbf{[5]}$$

Optimal Design and RE

For the three-level case, denote $\hat{\gamma}_{001} = \bar{Y}_T - \bar{Y}_C$ as the average treatment effect, where $\bar{Y}_T$ and $\bar{Y}_C$ are the mean outcome of the treatment and control conditions, respectively. For convenience, I assume the sample sizes are balanced within each level, hence, $n_{jk} = n$ and $J_k = J$. When treatment and control groups are the same size, the variance of the treatment effect estimate is

$$var(\hat{\gamma}_{001}) = \frac{2(Jn\rho_3 + n\rho_2 + \bar{\rho})\sigma_T^2}{KJn}, \quad \textbf{[6]}$$

where *n* is the number of Level 1 students in each Level 2 teacher, J is the number of Level 2 teachers in each Level 3 schools, and K is the number of Level 3 schools in both

the treatment and control group, ρ_2 is the Level 2 ICC, ρ_3 is the Level 3 ICC, $\bar{\rho} = 1 - \rho_2 - \rho_3$, and $\sigma_T^2 = \tau_\beta + \tau_\pi + \sigma^2$ is the total variance (Konstantopoulos, 2009). The total sample size is $2KJn$.

Optimal design. An optimal design specifies sample sizes for each level of nesting (i.e., n_{opt}, J_{opt}, and K_{opt}) for which the variance in the treatment effect estimate is minimized (Raudenbush, 1997), with respect to cost and other design parameters.[1] Typically, a linear cost function is used in the optimal design literature, though more complex functions are certainly possible if not likely. Equation [7] depicts a linear cost function for a three-level model,

$$2KJnC_1 + 2KJC_2 + 2KC_3 \leq C, \quad \textbf{[7]}$$

where *C* is the total budget, C_1 is the cost of an additional Level 1 student, and C_2 is the cost of an additional Level 2 teacher, and C_3 is the cost of an additional Level 3 school for either the treatment or control group.

According to Konstantopoulos (2008, 2009), optimal sample sizes for three-level models are as follows:

$$n_{opt} = \sqrt{\frac{C_2}{C_1}}\sqrt{\frac{(1-\rho_2-\rho_3)}{\rho_2}} \quad \textbf{[8]}$$

$$J_{opt} = \sqrt{\frac{C_3}{C_2}}\sqrt{\frac{\rho_2}{\rho_3}} \quad \textbf{[9]}$$

[1] Equivalently, an optimal design is achieved by maximizing the non-centrality parameter, λ, subject to the budget constraint.

$$K_{opt} = \frac{C}{2C_1 J_{opt} n_{opt} + 2C_2 J_{opt} + 2C_3} \quad \textbf{[10]}$$

An optimal design specifies an optimal allocation of resources in response to expectations in data variances in order to maximize the researcher's ability to detect a desired effect. Therefore, holding all else constant, an optimal design given one set of ICC values likely will be different than for an optimal design given a different set of ICC values. This fact is used to define an efficiency measure, RE, for three-level designs.

Relative efficiency. RE is defined as the ratio of the variance of the treatment effect estimate for a design with correctly specified ICCs to the variance of the treatment effect estimate for a design with incorrectly specified ICCs (Korendijk et al., 2010; Raudenbush, 1997). For a two-level model, Korendijk et al. (2010) presented RE in functional form as the $var(\hat{\gamma}_{01})^*$ for an optimal design given the true (population) ICC value, ρ^*, divided by the $var(\hat{\gamma}_{01})$ for an optimal design based on initial ICC estimate, ρ; hence,

$$RE = \frac{\left(\frac{n^*_{opt}\rho^* + \bar{\rho}^*}{J^*_{opt} n^*_{opt}}\right)}{\left(\frac{n_{opt}\rho^* + \bar{\rho}^*}{J_{opt} n_{opt}}\right)} = \frac{\left(\frac{n^*_{opt}\rho^* + \bar{\rho}^*}{\left(\frac{C}{2C_1 n^*_{opt} + 2C_2}\right) n^*_{opt}}\right)}{\left(\frac{n_{opt}\rho^* + \bar{\rho}^*}{\left(\frac{C}{2C_1 n_{opt} + 2C_2}\right) n_{opt}}\right)}. \quad \textbf{[11]}$$

In [11] [2], the designs (n^*_{opt}, J^*_{opt}) and (n_{opt}, J_{opt}) are optimal sample sizes derived by maximizing power given ρ^* and ρ, respectively, subject to the budget constraint $2JnC_1 + 2JC_2 \leq C$, where C is the total budget, C_1 is the cost of an additional Level 1 student, and C_2 is the cost of an additional Level 2 teacher for either the treatment or control group. In the third part of [11], the optimal number of Level 2 teachers has been written in terms of the budget constraint.

In the present study, I use RE to judge the robustness of designs from optimality as a result of misspecification of either ICC value. However, in order to explore the robustness of optimal designs for models involving three levels of nesting, [6] is first used to expand [11] to account for multiple ICCs.

In general, higher levels of RE are desirable. RE exists on the range (0, 1]; thus, to operationalize high levels of RE, a cut-off of $RE \geq 0.90$ is used. This value is consistent to the cut-off used by Korendijk et al. (2010). A value of $RE = 0.90$ suggests that an 11% (reciprocal of the RE) increase in sample size is needed to achieve a similar level of variance in the treatment effect estimate as a result of misspecification.

Throughout the analysis, surface plots and cross-sectional plots of RE for estimates of Level 2 and Level 3 ICCs are used to determine how RE is impacted according to changes in costs and true variances. To facilitate comparison of changes in costs and true variances, the range of misspecification of an ICC that maintains a high

---

[2] The reader should note that the variance of the treatment effect is always based on the true (observed) ICC, ρ^*, regardless of the initial ICC estimate. However, optimal sample sizes n^*_{opt}, J^*_{opt}, n_{opt}, and J_{opt} are based on true (starred) and initial (non-starred) ICC estimates, respectively. The estimate ρ does not appear in the equation because n_{opt} and J_{opt} are both functions of ρ.

level of RE is used. This range is referred to as the "range of acceptable misspecification." Alternatively, for directional comparisons, the terms *amount of acceptable under-specification* and *amount of acceptable over-specification* are used. The specific contexts (i.e., costs and true variances) considered are described in more detail next.

Assessment of the Robustness of the Optimal Design to Misspecification of One ICC

To gauge the robustness of the optimal design to misspecification of one ICC value, RE for a three-level model is plotted across estimates of both ρ_2 and ρ_3 relative to ρ_2^* and ρ_3^*, and subject to the constraints that $\bar{\rho}^* + \rho_2^* + \rho_3^* = 1$ and $\bar{\rho} + \rho_2 + \rho_3 = 1$. Using example ICCs from three-level models as a guide for ρ_2^* and ρ_3^* (see Table 2.1), I determine the ranges of acceptable misspecification, conditional on the correct specification of the other ICC. I consider combinations of ICC values less than or equal to 0.30 using increments of 0.05. For each case considered, costs for additional sample sizes are assumed to be the same to avoid distorting the graphs. I use $\frac{C_2}{C_1} = 5$ and $\frac{C_3}{C_2} = 5$, which are examples of commonly assumed values to explore the impacts of cost (Konstantopoulos, 2009; Raudenbush, 1997).

Specific relationships involving a combination of the two ICCs are also considered. For example, because the sum of the ICCs is bounded above, I consider instances where $\rho_2^* + \rho_3^*$ is large. Additionally, in each source listed in Table 2.1 above, examples existed where the Level 2 ICC was smaller than the Level 3 ICC as well as

where the Level 2 ICC was bigger than the Level 3 ICC. Thus, RE is explored for $\frac{\rho_2^*}{\rho_3^*} > 1$ as well as $\frac{\rho_2^*}{\rho_3^*} > 1$.

Next, the impact of varying costs is tested under fixed levels $\rho_2^* = 0.15$ and $\rho_3^* = 0.10$. These particular true ICC values are chosen to represent a typical educational design. Estimates of costs are also taken from the literature. Raudenbush (1997) presents costs for a 2-level model as a ratio of Level 2 to Level 1 costs, which range 2-50, and Konstantopoulos (2009) presents costs for a three-level model as a ratio of Level 3 to Level 2 costs on a range of 5-20. In this study I use cost ratios that range from 2-20 for $\frac{C_2}{C_1}$, and from 2-10 for $\frac{C_3}{C_2}$.

Additionally, general guidelines for maintaining high levels of RE with regard to misspecification of both ICCs are developed. The representative ICC values $\rho_2^* = .15$, and $\rho_3^* = .10$ as well as costs, $\frac{C_2}{C_1} = 5$ and $\frac{C_3}{C_2} = 5$, are again assumed to minimize the repetition of the analysis. Several important cases are considered. First, each ICC is considered conditionally on the misspecification of the other. For example, given over-specification of ρ_3 by 25% (i.e., $\rho_3 = 1.25\rho_3^*$), it is possible to find the range for ρ_2 that still maintains $RE \geq 0.90$. Examples with under-specification of each ICC by 20 and 40% as well as over-specification of each ICC by 20, 60 and 100% are presented.

Findings

Using [6] and [11], the RE for a 3-level model is derived as,

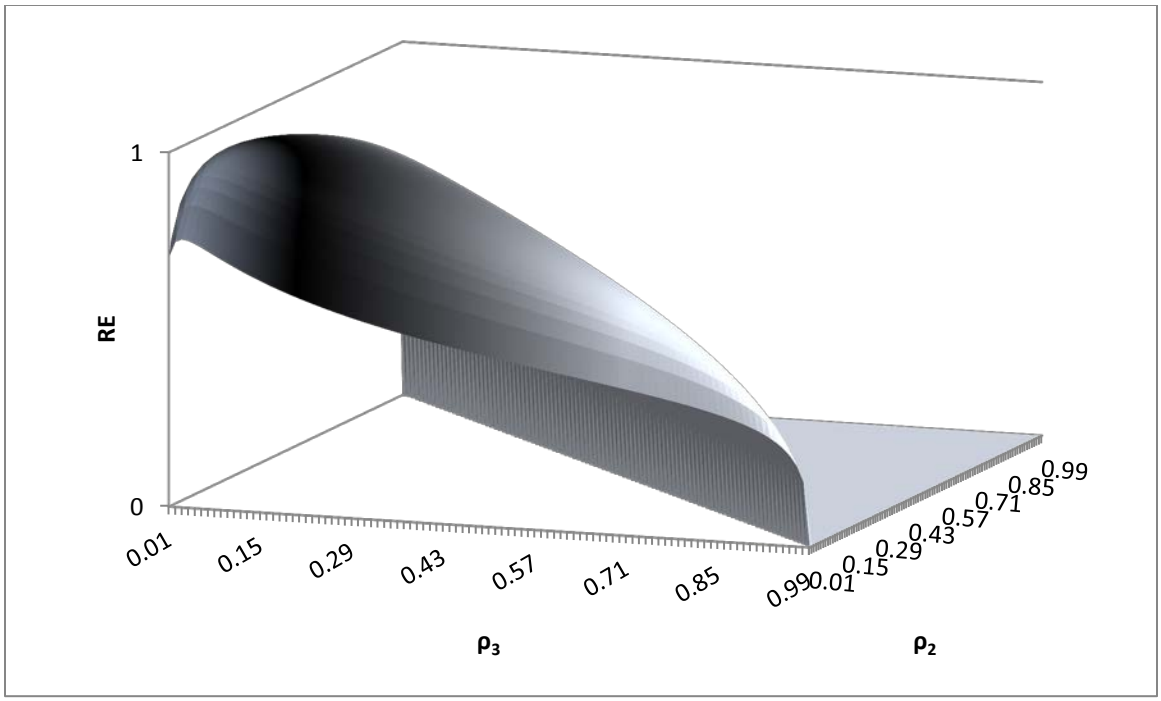
$$RE = \frac{\left(\frac{J^*_{opt}n^*_{opt}\rho^*_3 + n^*_{opt}\rho^*_2 + \bar{\rho}^*}{K^*_{opt}J^*_{opt}n^*_{opt}}\right)}{\left(\frac{J_{opt}n_{opt}\rho^*_3 + n_{opt}\rho^*_2 + \bar{\rho}^*}{K_{opt}J_{opt}n_{opt}}\right)} = \frac{\left(\frac{J^*_{opt}n^*_{opt}\rho^*_3 + n^*_{opt}\rho^*_2 + \bar{\rho}^*}{\left(\frac{\frac{C}{2}}{C_1J^*_{opt}n^*_{opt} + C_2J^*_{opt} + C_3}\right)J^*_{opt}n^*_{opt}}\right)}{\left(\frac{J_{opt}n_{opt}\rho^*_3 + n_{opt}\rho^*_2 + \bar{\rho}^*}{\left(\frac{\frac{C}{2}}{C_1J_{opt}n_{opt} + C_2J_{opt} + C_3}\right)J_{opt}n_{opt}}\right)}. \quad \textbf{[12]}$$

In [12], the designs ($n^*_{opt}, J^*_{opt}, K^*_{opt}$) and ($n_{opt}, J_{opt}, K_{opt}$) are optimal sample sizes derived by maximizing power given ρ^*_2 and ρ^*_3, or ρ_2 and ρ_3, respectively, subject to the budget constraint [7]. In the third part of [12], the optimal number of schools has been written in terms of the budget constraint. Furthermore, specific equations for optimal sample sizes were detailed above ([8] – [10]); using these equations, [12] can be rewritten as follows,

$$RE = \left(\frac{\sqrt{\left(\frac{C_3}{C_2}\right)\left(\frac{\rho_2^*}{\rho_3^*}\right)}\sqrt{\left(\frac{C_2}{C_1}\right)\left(\frac{\bar{\rho}^*}{\rho_2^*}\right)}\rho_3^* + \sqrt{\left(\frac{C_2}{C_1}\right)\left(\frac{\bar{\rho}^*}{\rho_2^*}\right)}\rho_2^* + \bar{\rho}^*}{\left(\frac{\frac{C}{2}}{C_1\sqrt{\left(\frac{C_3}{C_2}\right)\left(\frac{\rho_2^*}{\rho_3^*}\right)}\sqrt{\left(\frac{C_2}{C_1}\right)\left(\frac{\bar{\rho}^*}{\rho_2^*}\right)} + C_2\sqrt{\left(\frac{C_3}{C_2}\right)\left(\frac{\rho_2^*}{\rho_3^*}\right)} + C_3} \right)\sqrt{\left(\frac{C_3}{C_2}\right)\left(\frac{\rho_2^*}{\rho_3^*}\right)}\sqrt{\left(\frac{C_2}{C_1}\right)\left(\frac{\bar{\rho}^*}{\rho_2^*}\right)}} \right) \Bigg/ \left(\frac{\sqrt{\left(\frac{C_3}{C_2}\right)\left(\frac{\rho_2}{\rho_3}\right)}\sqrt{\left(\frac{C_2}{C_1}\right)\left(\frac{\bar{\rho}}{\rho_2}\right)}\rho_3^* + \sqrt{\left(\frac{C_2}{C_1}\right)\left(\frac{\bar{\rho}}{\rho_2}\right)}\rho_2^* + \bar{\rho}^*}{\left(\frac{\frac{C}{2}}{C_1\sqrt{\left(\frac{C_3}{C_2}\right)\left(\frac{\rho_2}{\rho_3}\right)}\sqrt{\left(\frac{C_2}{C_1}\right)\left(\frac{\bar{\rho}}{\rho_2}\right)} + C_2\sqrt{\left(\frac{C_3}{C_2}\right)\left(\frac{\rho_2}{\rho_3}\right)} + C_3} \right)\sqrt{\left(\frac{C_3}{C_2}\right)\left(\frac{\rho_2}{\rho_3}\right)}\sqrt{\left(\frac{C_2}{C_1}\right)\left(\frac{\bar{\rho}}{\rho_2}\right)}} \right) = \frac{\left(\frac{\sqrt{\left(\frac{C_3}{C_2}\right)\left(\frac{\rho_2^*}{\rho_3^*}\right)}\sqrt{\left(\frac{C_2}{C_1}\right)\left(\frac{\bar{\rho}^*}{\rho_2^*}\right)}\rho_3^* + \sqrt{\left(\frac{C_2}{C_1}\right)\left(\frac{\bar{\rho}^*}{\rho_2^*}\right)}\rho_2^* + \bar{\rho}^*}{\left(\frac{\frac{C}{2}}{C_1\sqrt{\left(\frac{C_3}{C_2}\right)\left(\frac{\rho_2^*}{\rho_3^*}\right)}\sqrt{\left(\frac{C_2}{C_1}\right)\left(\frac{\bar{\rho}^*}{\rho_2^*}\right)} + C_2\sqrt{\left(\frac{C_3}{C_2}\right)\left(\frac{\rho_2^*}{\rho_3^*}\right)} + C_3} \right)\sqrt{\left(\frac{C_3}{C_2}\right)\left(\frac{\rho_2^*}{\rho_3^*}\right)}\sqrt{\left(\frac{C_2}{C_1}\right)\left(\frac{\bar{\rho}^*}{\rho_2^*}\right)}} \right)}{\left(\frac{\sqrt{\left(\frac{C_3}{C_2}\right)\left(\frac{\rho_2}{\rho_3}\right)}\sqrt{\left(\frac{C_2}{C_1}\right)\left(\frac{\bar{\rho}}{\rho_2}\right)}\rho_3^* + \sqrt{\left(\frac{C_2}{C_1}\right)\left(\frac{\bar{\rho}}{\rho_2}\right)}\rho_2^* + \bar{\rho}^*}{\left(\frac{\frac{C}{2}}{C_1\sqrt{\left(\frac{C_3}{C_2}\right)\left(\frac{\rho_2}{\rho_3}\right)}\sqrt{\left(\frac{C_2}{C_1}\right)\left(\frac{\bar{\rho}}{\rho_2}\right)} + C_2\sqrt{\left(\frac{C_3}{C_2}\right)\left(\frac{\rho_2}{\rho_3}\right)} + C_3} \right)\sqrt{\left(\frac{C_3}{C_2}\right)\left(\frac{\rho_2}{\rho_3}\right)}\sqrt{\left(\frac{C_2}{C_1}\right)\left(\frac{\bar{\rho}}{\rho_2}\right)}} \right)}. \quad \textbf{[13]}$$

Figure 2.1 is a surface plot of [13], with $\frac{C_3}{C_2} = \frac{C_2}{C_1} = 5$, $\rho_2^* = .15$, and $\rho_3^* = .10$ which is used to illustrate the relationship between ρ_2 and ρ_3 in terms of RE. [3] The height of the surface is RE. RE reaches a maximum of 1 when both ICC estimates equal their true values. However, when either $\rho_2 \neq \rho_2^*$ or $\rho_3 \neq \rho_3^*$, the variance of the treatment effect estimate is not minimal in the denominator for [13], and therefore $RE < 1$.

[3] Notice that by standardizing the numerator and denominator by C_2, all references to cost can be written as ratios of costs between adjacent levels, which are assumed constants. Additionally, total cost is arbitrary, as it appears equally in the numerator and denominator.



Note. In this example, it is assumed $\rho_2^* = .15,\ \rho_3^* = .10,\ \frac{C_2}{C_1} = 5$, and $\frac{C_3}{C_2} = 5$.

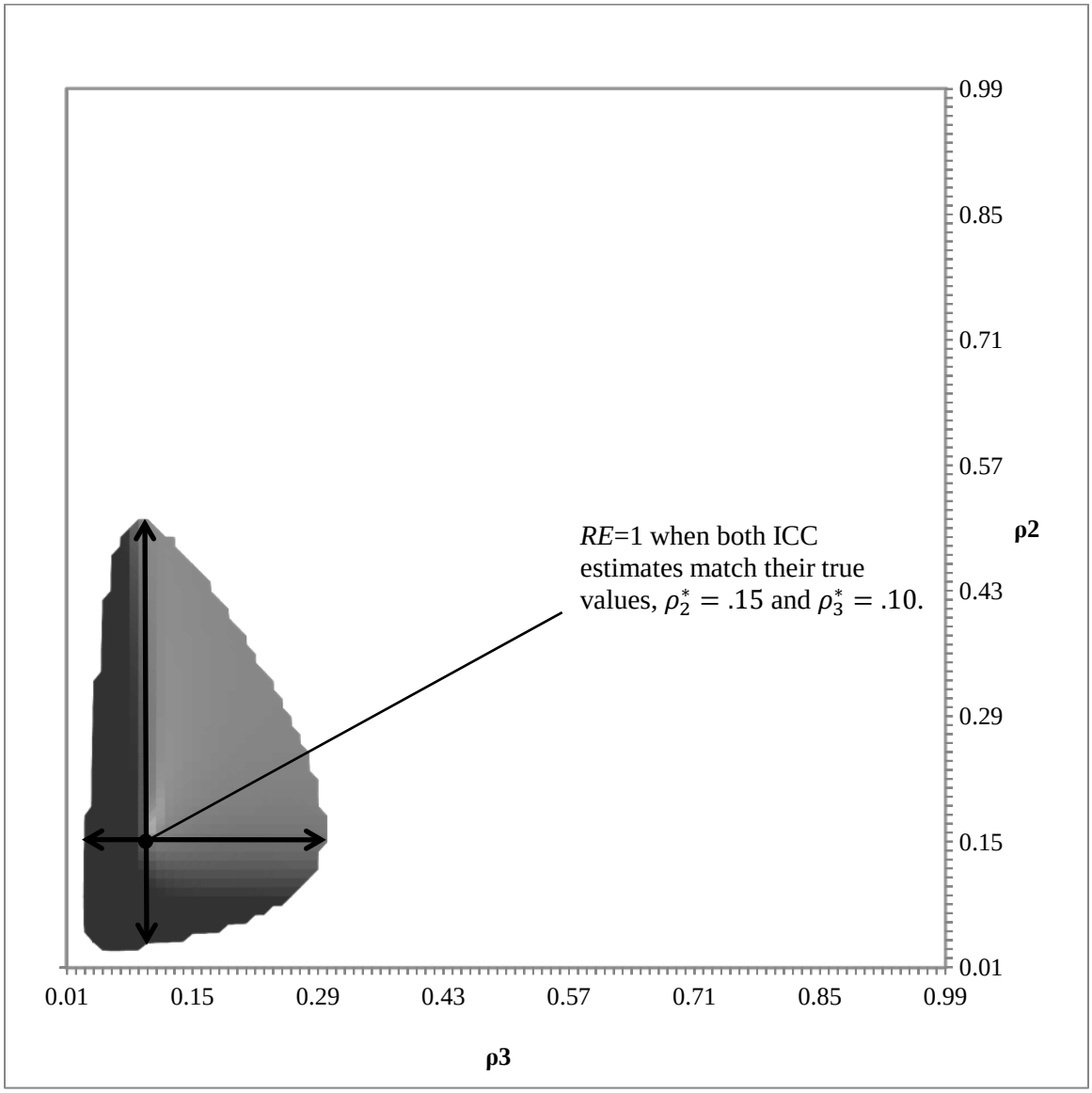
Figure 2.1. RE surface as a function of ρ_2 and ρ_3.

It is clear from [13] the range of misspecification for ICCs that maintains an acceptable range of RE is impacted by context (i.e., costs and true variances). Below, several contextual factors that influence RE are explored: correct specification of one ICC, important relationships involving both ICCs, relative costs of additional participants, and simultaneous misspecification of both ICCs.

Correct Specification of One ICC

As shown in Figure 2.1, RE declines more rapidly for under-specification of ρ_2^* and ρ_3^* than for over-specification. This is more easily illustrated through a contour map. Figure 2.2 illustrates the same example, only in two dimensions, with RE represented by

shading. In this plot, only values with $RE \geq 0.90$ are shown. The maximum RE is highlighted at the true ICC values.



Note. In this example, it is assumed $\rho_2^* = .15$, $\rho_3^* = .10$, $\frac{C_2}{C_1} = 5$, and $\frac{C_3}{C_2} = 5$.

Figure 2.2. RE contour map as a function of ρ_2 and ρ_3 for $RE \geq 0.90$.

When $\rho_3 = \rho_3^*$, the range of acceptable misspecification of ρ_2^* is shown using a vertical line. This line can be divided into over-specification (above the true value), and under-specification (below the true value). When $\rho_2 = \rho_2^*$, the range of acceptable misspecification of ρ_3^* is represented by a horizontal line. This line can also be divided into over-specification (right of the true value), and under-specification (left of the true value). Clearly there is more shaded area above (representing over-specification of ρ_2^*) and to the right (representing over-specification of ρ_3^*) of the true value, than below or to the left (representing under-specification of the respective values).

The empirical literature (see Table 2.1) suggests that ρ_2^* and ρ_3^* typically range from 0.05 to 0.30. Thus, using an increment of 0.05, the range of acceptable misspecification is explored for combinations of ρ_2^* and ρ_3^*.

Table 2.2 shows the amount of acceptable under-specification and over-specification for the various combinations of ρ_2^* and ρ_3^*; several conclusions are apparent. First, for estimates ρ_2 and ρ_3, the range of acceptable misspecification is more influenced by its corresponding true value, ρ_2^* and ρ_3^*, respectively. Next, the acceptable amount of under-specification for the ICCs does not vary considerably for different ρ_2^* and ρ_3^*. $RE \geq 0.90$ is maintained for under-specification of both ICC values by up to 60% for smaller values of ρ_2^* and ρ_3^*, and by up to 80% and 70% for larger values of ρ_2^* and ρ_3^*, respectively. Finally, across the range of true values considered, there is significantly more variance in the amount of acceptable over-specification with regard to ρ_2^* than ρ_3^* . Depending on ρ_2^* and ρ_3^*, the amount of acceptable over-specification in the estimate ρ_2 may be as little as 100% or as much as 480%, while the amount of acceptable over-

specification in the estimate ρ_3 may be as little as 93% or as much as 240%. In general, for larger ρ_2^* and ρ_3^*, the amount of acceptable over-specification for either of the estimate is smaller.

Table 2.2

Ranges of Acceptable Misspecification for Estimates ρ_2 and ρ_3

ρ_2

Amount of acceptable under-specification (%)							Amount of acceptable over-specification (%)						
	ρ_3^*							ρ_3^*					
ρ_2^*	0.05	0.10	0.15	0.20	0.25	0.30	ρ_2^*	0.05	0.10	0.15	0.20	0.25	0.30
0.05	80	80	80	80	80	80	0.05	340	380	420	440	460	480
0.10	70	70	70	80	80	80	0.10	260	280	290	300	300	310
0.15	73	73	73	73	73	73	0.15	207	220	227	227	227	220
0.20	70	70	75	75	75	75	0.20	170	180	180	180	175	165
0.25	68	72	72	72	72	72	0.25	144	148	148	144	136	128
0.30	70	70	70	70	73	73	0.30	123	123	120	117	110	100

ρ_3

Amount of acceptable under-specification (%)							Amount of acceptable over-specification (%)						
	ρ_3^*							ρ_3^*					
ρ_2^*	0.05	0.10	0.15	0.20	0.25	0.30	ρ_2^*	0.05	0.10	0.15	0.20	0.25	0.30
0.05	60	70	67	70	68	70	0.05	220	200	180	155	136	120
0.10	60	70	67	70	68	67	0.10	240	200	180	155	136	117
0.15	60	70	67	70	68	67	0.15	240	200	173	150	132	113
0.20	60	70	67	70	68	67	0.20	240	200	173	150	128	107
0.25	60	70	67	70	68	67	0.25	240	200	173	145	120	100
0.30	60	70	67	70	68	67	0.30	240	200	167	140	116	93

Important Relationships Involving Both ICCs

Since the true ICC values ρ_2^* and ρ_3^* are significant drivers of the shape of the RE surface, certain relationships involving both of these values also warrant exploration. Below, two important cases are considered: $\rho_2^* + \rho_3^*$ large and $\frac{\rho_2^*}{\rho_3^*} <> 1$.

First, because ICCs are percentages, $\rho_2^* + \rho_3^*$ must be greater than 0, and cannot sum to more than 1. In the present study, actual estimates of educational ICC values from the literature have motivated examples with smaller true ICC values. However, as the sum of ρ_2^* and ρ_3^* gets larger, due to a ceiling effect, RE will decline more rapidly for over-specification of ρ_2 and ρ_3 than for under-specification.

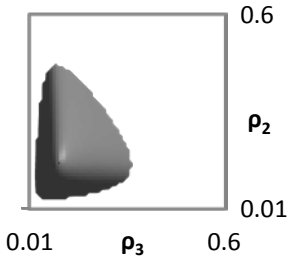
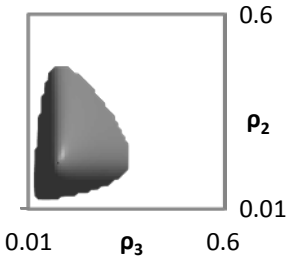
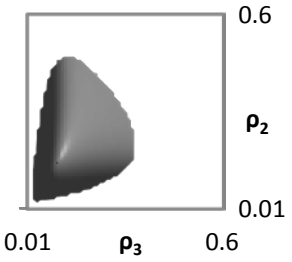
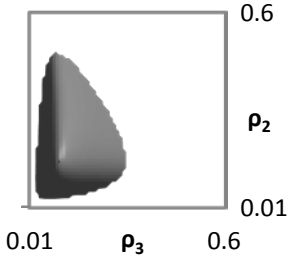
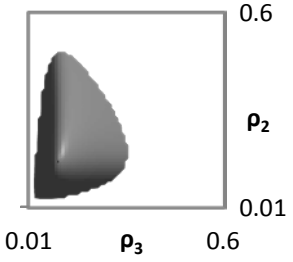
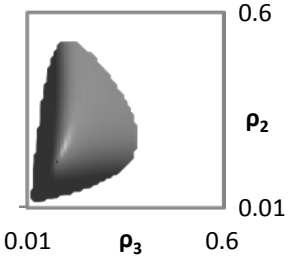
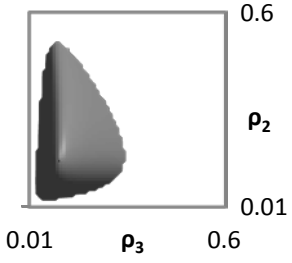
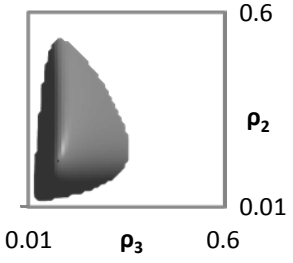
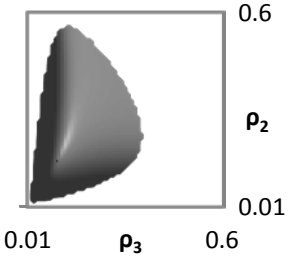
Another important relationship to consider is $\frac{\rho_2^*}{\rho_3^*}$ (where the true ICCs are small, as is typically true in education.) As a benchmark, consider the case when $\frac{\rho_2^*}{\rho_3^*} = 1$, or equivalently, when $\rho_2^* = \rho_3^*$ Both the amount of acceptable under-specification and the amount of acceptable over-specification are larger for ρ_2^* than for ρ_3^* (see Table 2.2); however, the differences in these amounts for estimates ρ_2 and ρ_3 vary according to the size of the true values. For larger true values (e.g., $\rho_2^* = \rho_3^* = 0.30$), the ratio between the amount of acceptable under-specification and over-specification of estimates ρ_2 and ρ_3 is small, approximately 1.08 for under-specification and 1.09 for over-specification. For smaller true values (e.g., $\rho_2^* = \rho_3^* = 0.05$), the ratios are bigger, approximately 1.33 for under-specification and 1.55 for over-specification.

Now consider the case when $\frac{\rho_2^*}{\rho_3^*} < 1$. For example, let $\rho_2^* = 0.10$ and $\rho_3^* = 0.25$. A similar result to the benchmark example is found. The amount of acceptable under-specification as well as the amount of acceptable over-specification are larger when estimating ρ_2 than when estimating ρ_3. The ratio in the amount of acceptable under-specification is 1.18, while the ratio in the amount of acceptable over-specification is 2.21.

For the case when $\frac{\rho_2^*}{\rho_3^*} > 1$, a different result is found. Consider a situation with $\rho_2^* = 0.25$ and $\rho_3^* = 0.10$. Now, the amount of acceptable under-specification is larger when estimating ρ_2 than when estimating ρ_3, but the amount of acceptable over-specification is smaller when estimating ρ_2 than when estimating ρ_3. The ratio in the amount of acceptable under-specification is 1.03, while the ratio in the amount of acceptable over-specification is 0.74.

Relative Costs of Additional Participants

Figure 2.3 shows how RE is impacted for specific ICCs under varying cost structures, assuming $\rho_2^* = 0.15$ and $\rho_3^* = 0.10$. Within each row, the cost for an additional Level 2 teacher relative to the cost of a Level 1 student increases from left to right. Similarly, within each column, the cost for an additional Level 3 school relative to the cost of a Level 2 teacher increases from top to bottom.

	$\frac{C_2}{C_1} = 2$	$\frac{C_2}{C_1} = 5$	$\frac{C_2}{C_1} = 20$
$\frac{C_3}{C_2} = 2$	A 	B 	C 
$\frac{C_3}{C_2} = 5$	D 	E 	F 
$\frac{C_3}{C_2} = 10$	G 	H 	I 

Note. In this example, it is assumed $\rho_2^* = .15$ and $\rho_3^* = .10$.

Figure 2.3. RE contour maps under various cost ratios, $\frac{C_2}{C_1}$ and $\frac{C_3}{C_2}$.

Comparing the different panels in Figure 2.3, it is clear that changes in the relative cost of additional participants across levels impacts the acceptable range of misspecification of ICCs. However, the overall impact is relatively small. This is illustrated by comparing the examples depicted in panels A and I. The acceptable amount of under-specification of ρ_2^* and ρ_3^* does not change; it is 73% and 70%, respectively, in both examples. The acceptable amount of over-specification of ρ_2^* grows from 193% for the example in panel A to 220% for the example in panel I, while the acceptable amount of over-specification of ρ_3^* shrinks from 260% for the example in panel A to 230% for the example in panel I. Neither a ten-fold increase in the cost of an additional Level 2 teacher relative to the cost of a Level 1 student, nor a five-fold increase in the cost of an additional Level 3 school relative to the cost of a Level 2 teacher changes the amount of acceptable under-specification of ρ_2^* and ρ_3^* by much. These large changes in cost have at most a 14% impact the amount of acceptable over-specification of ρ_2^* and ρ_3^*.

Simultaneous Misspecification of Both ICCs

Earlier, the acceptable range of misspecification for each ICC was considered conditionally on the correct specification of the other. In this section the range of acceptable misspecification for each ICC is considered conditionally on the misspecification of the other.

Table 2.3 presents the amount of over-specification and under-specification for each ICC estimate conditional on over-specification and under-specification of the other.

For convenience, it is again assumed that $\rho_2^* = .15$, and $\rho_3^* = .10$, while cost ratios are also held constant with $\frac{C_2}{C_1} = 5$ and $\frac{C_3}{C_2} = 5$. The examples considered include 20 and 40% under-specification and 20, 60 and 100% over-specification.

Table 2.3

Respective Ranges of Acceptable Misspecification for Estimates ρ_2 and ρ_3 Conditional on Misspecification of ρ_3^* and ρ_2^*

Acceptable Range of Misspecification for $\rho_3|\rho_2$

		Range of misspecification for ρ_3 where $RE \geq 0.90$		Amount of under-misspecification and over-misspecification of ρ_3 where $RE \geq 0.90$	
ρ_2	Amount of Misspecification	Lower Bound	Upper Bound	Under-Specification	Over-Specification
0.09	40% under	0.03	0.26	70%	160%
0.12	20% under	0.03	0.29	70%	190%
0.15	None	0.03	0.30	70%	200%
0.18	20% over	0.04	0.31	60%	210%
0.24	60% over	0.04	0.29	60%	190%
0.30	100% over	0.04	0.28	60%	180%

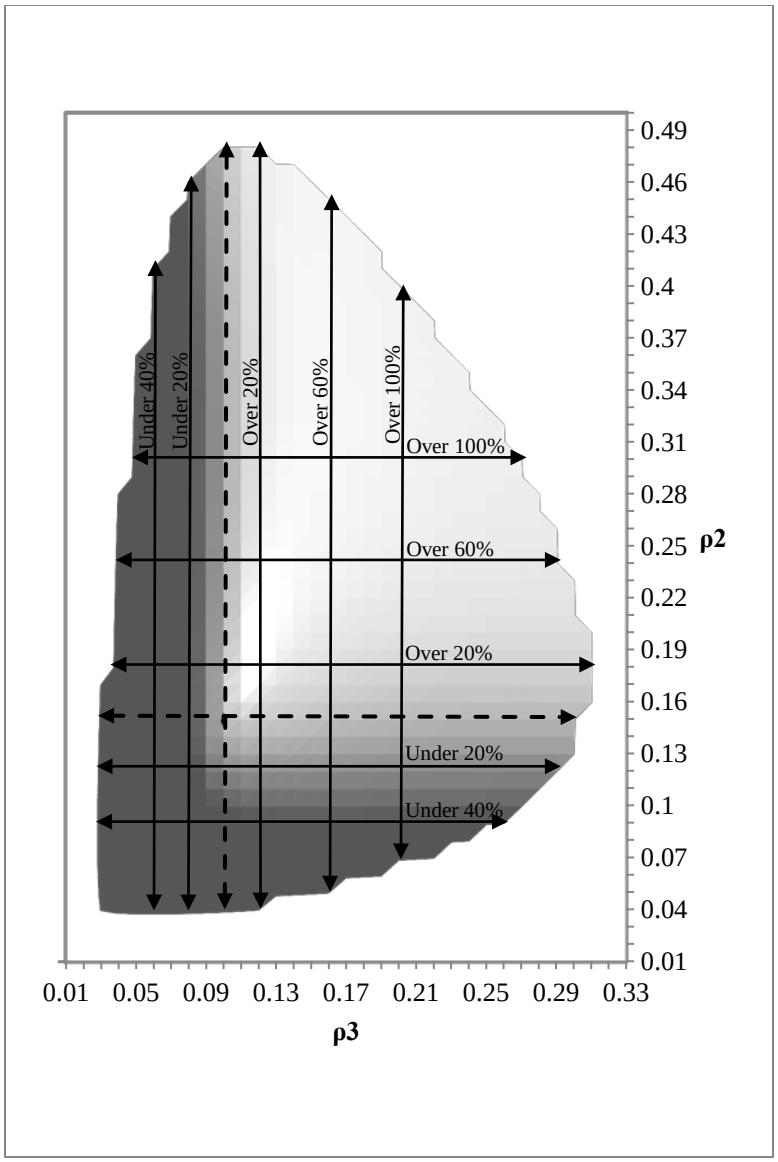
Acceptable Range of Misspecification for $\rho_2|\rho_3$

		Range of misspecification for ρ_2 where $RE \geq 0.90$		Amount of under-misspecification and over-misspecification of ρ_2 where $RE \geq 0.90$	
ρ_3	Amount of Misspecification	Lower Bound	Upper Bound	Under-Specification	Over-Specification
0.06	40% under	0.04	0.41	73%	173%
0.08	20% under	0.04	0.46	73%	207%
0.10	None	0.04	0.48	73%	220%
0.12	20% over	0.04	0.48	73%	220%
0.16	60% over	0.05	0.45	67%	200%
0.20	100% over	0.07	0.40	53%	167%

Note. In this example, it is assumed $\rho_2^* = .15$, $\rho_3^* = .10$, $\frac{C_2}{C_1} = 5$, and $\frac{C_3}{C_2} = 5$.

For both ρ_2 and ρ_3, the range of misspecification does not change much for small amounts of conditional misspecification of the other ICC. For under-specification and over-specification of ρ_2^* and ρ_3^* by 20%, the respective amount of acceptable under-specification and over-specification of the other ICC is nearly identical to when specification is correct. However, when misspecification is large for one ICC, the range of acceptable misspecification for the other ICC can be dramatically reduced. This is particularly true for under-specification more extreme than 40% and over-specification more extreme than 100%.

Figure 2.4 depicts the example above to illustrate how the range of acceptable misspecification changes for various combinations of conditional under-specification and conditional over-specification. When under-specification and over-specification of one ICC becomes extreme, the slope of the boundary of RE is steep relative to the corresponding ICC axis, and small amounts of additional misspecification for one ICC can produce a large change in the range of acceptable misspecification for the other ICC. In contrast, when misspecification is minimal, the slope of the boundary where $RE = 0.90$ is flat relative to the corresponding ICC axis, and small amounts of additional misspecification for one ICC do not produce a significant change in the range of acceptable misspecification for the other ICC.



Note. In this example, it is assumed $\rho_2^* = .15,\ \rho_3^* = .10,\ \frac{c_2}{c_1} = 5$, and $\frac{c_3}{c_2} = 5$.

Figure 2.4. Graphical depiction of the respective ranges of acceptable misspecification for estimates ρ_2 and ρ_3 conditional on misspecification of ρ_3^* and ρ_2^*.

Discussion

As evaluators begin to appropriately account for additional levels of nesting with larger and more sophisticated CRTs, the need to understand the implications of parameter

misspecifications is heightened. Few empirical estimates exist of parameter values for three-level models, and therefore CRTs with three levels are prone to misspecification. When misspecification occurs, the resulting CRT can be underpowered and effectively useless, or overpowered and waste important resources. In a discipline like education where resources for evaluation are scarce, properly powering a design is essential practice. Furthermore, according to established standards, evaluators are charged with using cost-effective methods (Yarbrough, Shulha, Hopson, & Caruthers, 2011), and by definition, a relatively inefficient design is not cost-effective.

Given the shape of the RE curve in two-dimensions (Korendijk et al., 2010), the resulting shape of the RE surface in three dimensions might be expected; a similar shape can be seen in the profile along each axis. Grounding the analysis to correspond to empirical estimates of ICCs in education, this study concludes that for both ρ_2^* and ρ_3^*, the amount of acceptable over-specification is much larger than the amount of acceptable under-specification. For example, for representative true ICC values $\rho_2^* = 0.15$ and $\rho_3^* = 0.10$ (assuming $\frac{C_2}{C_1} = 5$ and $\frac{C_3}{C_2} = 5$), the range of acceptable misspecification for estimate ρ_2 is from 73% under-specified to 220% over-specified. For estimate ρ_3 this range is from 70% under-specified to 200% over-specified. Additionally, in this study it has be shown that when true ICC values are equivalent, the range of acceptable misspecification is larger for estimate ρ_2 than for estimate ρ_3.

In this study, small fluctuations in costs or true ICC values appear to have little impact on the range of misspecification that maintains a high level of RE. Hence, in most situations optimal designs are quite robust to misspecification of an ICC. However, when

costs or true ICC values are extreme, the acceptable range of misspecification of either ICC value can shrink dramatically.

Careful inspection of Figure 2.4 reveals that the range of acceptable misspecification for estimates ρ_2 and ρ_3 is not necessarily maximal where $\rho_2 = \rho_2^* = 0.15$ and $\rho_3 = \rho_3^* = 0.10$. Rather, in this case, slight over-specification of ρ_2^* yields a wider range of acceptable misspecification of ρ_3. Such a fact could be useful for evaluators estimating variances in situations where variance at one level is believed to be larger than variance at the other level. However, as the shape of the contour map can change substantially depending on the values of the true ICCs, this is not always the case.

Evaluators designing CRTs with three-levels in education face the challenge of not having precise estimates of variances across the different levels of nesting. However, the results of this study suggest the additional level of nesting provides increased cushioning for maintaining a high level of RE if one of the two estimates is accurate. Korendijk et al. (2010) suggest an acceptable range of misspecification is defined by 75% under-specification and 175% over-specification. Using a representative example with $\rho_2^* = 0.15$, $\rho_3^* = 0.10$, $\frac{C_2}{C_1} = 5$ and $\frac{C_3}{C_2} = 5$, conditional on the correct specification of $\rho_2 = \rho_2^*$, the amount of acceptable under-specification for estimate ρ_3 is a bit smaller, 70%, than the cut-off proposed for a two-level design, but the amount of acceptable over-specification for estimate ρ_3, 200%, is 14% larger. Likewise, using the same example, conditional on the correct specification of $\rho_3 = \rho_3^*$, the amount of acceptable under-specification for estimate ρ_2 is only slightly smaller, 73%, than the cut-off proposed for a

two-level design, but the amount of acceptable over-specification for estimate ρ_2, 220%, is 26% bigger.

Unfortunately, misspecification is likely to occur on both ICCs, and this negatively impacts the range of acceptable misspecification for the ICCs. However, for the same representative sample, this study demonstrates that only when one ICC is under-specified by more than 40% or over-specified by more than 100% does the range of under-specification and over-specification for the other ICC drop to the levels found in the two-level model.

Application

To illustrate the findings of this study, consider the following example. Suppose an evaluator is asked to assess the impact of a new biology curriculum by a state agency. The new biology curriculum will be implemented for tenth graders in schools across the state. The evaluator is concerned about the threat of contamination and therefore plans to use a three-level CRT with students nested in teachers nested in schools for the curriculum study. The outcome measure will be student scores on a standardized science test for grade 10.

To design the study the evaluator needs to conduct a power analysis by estimating various design parameter values. Suppose the evaluator wants the study to have power of at least 0.80, and be able to detect an effect of at least 0.20 with 95% confidence. For simplicity of explanation, assume that the evaluator does not have access to a covariate to

explain variance.[4] Assume the cost structure for implementing the CRT has $\frac{C_2}{C_1} = 5$ and $\frac{C_3}{C_2} = 5$. The evaluator next must estimate the percentage of total variance that occurs among teachers and among schools.

Assume the ICC estimates used by the evaluator for this hypothetical study are taken from Zhu, Jacob, Bloom, and Xu (2012), who used North Carolina end-of-course data to develop empirical estimates of ICCs for studies of biology. According to Zhu et al. (2012), an estimate of the school-level (Level 3) ICC is 0.077 and teacher-level (Level 2) ICC is 0.293. Using equations [8] - [10], the evaluator will estimate the optimal sample sizes[5] for students, teachers, and schools as:

$$n_{opt} = \sqrt{5}\sqrt{\frac{(1 - 0.293 - 0.077)}{0.293}} = 3.279 \quad \textbf{[14]}$$

$$J_{opt} = \sqrt{5}\sqrt{\frac{0.293}{0.077}} = 4.362 \quad \textbf{[15]}$$

$$K_{opt} = \frac{\frac{C}{2C_2}}{0.2 * 3.279 * 4.362 + 4.362 + 5} = \frac{C}{24.445C_2}. \quad \textbf{[16]}$$

---

[4] Zhu, Jacob, Bloom, and Xu (2012) provide estimates of R^2 values for a school-level and student-level covariates using end-of-course assessments from North Carolina from 2005. Their estimates of variance explained by a school-level covariate are 0.675 at the school-level, 0.003 at the classroom-level, and 0.000 at the student-level. Alternatively, they estimate that a student-level covariate could explain 0.229, 0.693, and 0.310% of the variance, respectively, at the school-, classroom-, and student-level. Thus, use of a covariate would significantly reduce the number of schools needed in the study.

[5] Here I assume that fractional units are acceptable for demonstrative purposes. When assuming a balanced design, fractional units are theoretically impossible in an educational setting. However, if the requirement of balance is relaxed, then the harmonic mean number of units is typically used as the sample size measure, and may be fractional.

Clearly the optimal design depends on the budget. With a larger budget, the evaluator can include more schools. Referring to Konstantopoulos (2009), for the desired power level of 0.80 with significance level 0.05, the design will need 149.68 schools, and hence a total of 4,281.39 participants in order to detect the desired effect. The evaluation will proceed with the experiment under sample sizes ([14] - [16]), and when the experiment has concluded the evaluator will learn the true ICCs (recall that no covariates are used in this example.)

Suppose the true ICCs are $\rho_2^* = 0.15$, $\rho_3^* = 0.10$. These values mean the original estimates represent under-specification of the true Level 3 ICC by approximately 20% and over-specification of the true Level 2 ICC by approximately 100%. At first glance, the North Carolina estimates do not appear to translate well for this particular study. Using these estimates for the ICCs, misspecification will occur in both ICCs. However, according to the analysis (see Table 2.3), misspecification, even at these levels, will still yield a reasonably high level of RE. This can also be seen by looking at Figure 2.4, where the intersection of the vertical "20% Under" line and the horizontal "100% Over" line is still in the shaded region. The optimal sample sizes given the true variance and cost structure are

$$n_{opt}^* = \sqrt{5}\sqrt{\frac{(1 - 0.15 - 0.10)}{0.15}} = 5 \quad \textbf{[17]}$$

$$J_{opt}^* = \sqrt{5}\sqrt{\frac{0.15}{0.10}} = 2.739 \quad \textbf{[18]}$$

$$K_{opt}^{*} = \frac{\frac{C}{2 * C_2}}{0.2 * 5 * 2.739 + 2.739 + 5} = \frac{C}{20.954C_2}. \quad \textbf{[19]}$$

For the desired power level of 0.80 to detect an effect of size 0.20, with significance level 0.05, the design will need 166.41 schools, and therefore a total of 4,557.36 participants.

If the true ICCs are as described above, RE will be less than one because the optimal sample sizes are determined by the budget, which in this example is dictated by [16]. To illustrate this point, suppose $C_2 = 1$. Then, for the model with incorrectly specified ICCs, *C*=3,658.83 in order to achieve power of 0.80, but for the model with correctly specified ICCs, only *C*=3,487.06 was needed to achieve power of 0.80. The difference in budget suggests inefficiency. In effect, more resources to achieve the desired power level were used than needed because sample size estimates for the true case are derived using the wrong ICC values. Under correctly specified ICCs, the additional resources could have been allocated optimally, which would result in no change to equations [14] and [15] (these equations are based on the cost ratios between levels and the ICCs which are constant in the two cases), but equation [16] will instead be

$$K_{opt} = \frac{\frac{3{,}658.83}{2 * 1}}{0.2 * 5 * 2.739 + 2.739 + 5} = 174.61. \quad \textbf{[20]}$$

Then [17], [18], and [20], represent the correct specification with the true ICCs given the budget dictated by the original estimates, and the variance of the treatment effect estimate for this model is 0.0024. The variance of the treatment effect estimate for [14], [15], and [16] and the true ICCs is 0.002496. Thus, *RE*=0.9614, suggesting that a (1/.9614=1.04)

4% larger sample size is needed in order to achieve the same level in the variance of the treatment effect estimate.

Closing Remarks

Although it appears there is a considerable amount of flexibility for evaluators to estimate ICC values for a three-level CRT, and more flexibility than in a two-level model, evaluators should still strive to properly estimate ICCs by developing more rigorous empirical estimates. Misspecification of parameter values, even when a reasonably high level of RE is maintained, still is an inefficient design. When CRTs are efficiently designed, resources are saved. The high cost of conducting a CRT means that misspecification of design parameter values can translate into significant waste, through over-powering or under-powering a study. With efficient designs, these resources can be directed toward more educational programming, or other deserving research and evaluation efforts. It is important that designs are powered precisely so that research dollars are maximized to grow the evidence-base.

References

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. (in press). Intraclass correlations and covariate outcome correlations for planning 2 and 3 level cluster randomized experiments in education. *Evaluation Review*.

Konstantopoulos, S. (2008). The power of the test for treatment effects in three level cluster randomized designs. *Journal of Research on Educational Effectiveness, 1*(2), 66-88. doi:10.1080/19345740701692522

Konstantopoulos, S. (2009). Incorporating cost in power analysis for three-level cluster-randomized designs. *Evaluation Review, 33*(4), 335-357. doi:10.1177/0193841X09337991

Korendijk, E. J., Moerbeek, M., & Maas, C. J. (2010). The robustness of designs for trials with nested data against incorrect initial intracluster correlation coefficient estimates. *Journal of Educational and Behavioral Statistics, 35*(5), 566-585. doi:10.3102/1076998609360774

Raudenbush, S. W. (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods, 2*(2), 173-185. doi:10.1037/1082-989X.2.2.173

Spybrook, J. K. (in press). Detecting intervention effects across context: An examination of the precision of cluster randomized trials. *Journal of Experimental Education*. doi:10.1080/00220973.2013.813364

Spybrook, J. K., & 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

Westine, C. D., Spybrook, J. K., & Taylor, J. A. (in press). An empirical investigation of variance design parameters for planning cluster-randomized trials of science achievement. *Evaluation Review*.

Yarbrough, D. B., Shulha, L. M., Hopson, R. K., & Caruthers, F. A. (2011). *The program evaluation standards: A guide for evaluators and evaluation users* (3rd ed.). Thousand Oaks, CA: Sage.

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

CHAPTER III

AN EMPIRICAL INVESTIGATION OF VARIANCE DESIGN PARAMETERS FOR PLANNING CLUSTER-RANDOMIZED TRIALS OF SCIENCE ACHIEVEMENT[6]

In the past decade, there has been a dramatic shift in educational policy towards the use of randomized trials (RTs) to test the effectiveness of educational programs, policies, and practices. Given the natural clusters in the U.S. education system involving students nested within classrooms, classrooms nested within schools, and schools nested within districts, and the fact that interventions are typically administered at the classroom-, school-, or district-level, a specific category of RTs, cluster-randomized trials (CRTs), are common (Bloom, 2005; Boruch & Foley, 2000; Cook, 2005). CRTs rely on random assignment of intact clusters to treatment conditions, such as the classroom or school (Raudenbush & Bryk, 2002).

The rise of CRTs to determine the effectiveness of educational interventions is clear from funding trends by agencies such as the Institute of Education Sciences (IES). Since 2002, the National Center for Education Research (NCER) within the IES has funded more than 100 CRTs (Institute of Education Sciences, 2013b; Spybrook & Raudenbush, 2009). Compare this to the few CRTs funded prior to 2002 by the Department of Education and the shift is overwhelming (Mosteller & Boruch, 2002). The

[6] The most recent version of this chapter has been accepted for publication with the peer-reviewed journal *Evaluation Review* by SAGE.

majority of these CRTs focus on examining the effectiveness of reading and mathematics programs and practices. The studies focus on all grade levels ranging from pre-K through high school, with the majority of them focusing on pre-K and the elementary grades. We have also started to see more CRTs of interventions to boost science achievement from major grant funders. For example, the What Works Clearinghouse lists 13 CRTs for science achievement started since 2005 in its registry of RTs (Institute of Education Sciences, 2013a).

In order for CRTs to yield high-quality evidence of whether a program is effective, among other things, such studies must be well-designed with adequate power to detect a treatment effect of a reasonable magnitude. The field has made substantial progress in terms of how to calculate statistical power for CRTs (Donner & Klar, 2000; Raudenbush, 1997; Raudenbush & Liu, 2000; Raudenbush, Martinez, & Spybrook, 2007; Schochet, 2008). One of the key concepts emerging from this line of investigation is the importance of good design parameters to use in the power analyses, noting that the power analysis is only as accurate as the design parameters. That is, if any of the design parameters are inaccurate, then too many or too few schools may be recruited resulting in unnecessary costs or an underpowered trial.

As a result, there has been a growing body of literature in the past several years providing empirical estimates of design parameters necessary for statistical power calculations for CRTs in education. Aside from meta-analytical work to estimate effect sizes, and recent work by Kelcey and Phelps (2013), which expands the discussion to teacher-level outcomes, the literature on empirical estimates of design parameters has

largely revolved around using student outcomes to estimate intraclass correlations (ICCs) and percent of variance explained (R^2) by pretest and demographic covariates. Early endeavors focused on estimating these parameters for two level models using national longitudinal survey data and data from individual districts or program evaluations (Bloom, Bos, & Lee, 1999; Bloom, Richburg-Hayes, & Black, 2007; Hedges & Hedberg, 2007). More recent work has utilized entire state databases of student achievement as well as program evaluation data to estimate ICCs and R^2 values for models with three levels (Hedberg & Hedges, 2011; Jacob, Zhu, & Bloom, 2010; Konstantopoulos, 2009; Zhu, Jacob, Bloom, & Xu, 2012).

The majority of this literature has emphasized empirical estimates for elementary, middle, and high school students and is predominately focused on reading and mathematics outcomes. Zhu et al. (2012) briefly examined design parameters for science outcomes. However, this study was limited to end-of-class tests for specific science classes, i.e., biology, chemistry, and was restricted to high school grades. We are unaware of any systematic investigations into design parameters for science outcomes across elementary, middle, and high school grades. In addition, we are unaware of any estimates that use a standardized science test as the outcome, which is a common outcome in intervention studies. The lack of empirical estimates of design parameters for science outcomes makes it challenging for researchers to design CRTs to test science interventions. Researchers are often forced into using empirical estimates for mathematics and reading design parameters since none exist for science outcomes.

However, it is unclear whether this borrowing of empirical estimates of design parameters is appropriate.

Challenges with Borrowing ICCs

Currently we lack empirical estimates of ICCs for science outcomes to know whether reading and mathematics ICCs are a good proxy for science. To further complicate things, estimates of reading and mathematics ICCs are often inconsistent. Drawing from the literature of available ICC estimates, Table 3.1 provides a summary of the unconditional ICCs for grades 5, 8, 10, and 11 for reading and mathematics (ICCs for reading in grade 11 are not reported by any source) for models with two levels. We specifically report for grades 5, 8, 10, and 11 because unlike reading and mathematics, science is not tested annually. In fact, only 3 states tested science annually in grades 3 through 12 whereas 18 states tested science in grades 5, 8, 10, and 11 (Time4Learning, 2013). Of the remaining states, a combination of grades 5, 8, 10, and 11 are the most commonly tested. From Table 3.1 it is clear that in many cases there is inconsistency in the unconditional ICCs for reading and mathematics. For example, consider grade 10 for the study conducted by Hedges and Hedberg (2007). They found an ICC in grade 10 of 0.234 for mathematics and 0.183 for reading. These two values are quite different and would lead to different power calculations. For example, suppose a researcher believes the science ICC is likely to be similar to mathematics and chooses to borrow this value. If in fact the ICC is more similar to reading, the researcher overestimated the ICC initially and may have an overpowered study. If the researcher borrowed from reading when in fact the ICC is more similar to mathematics, the initial ICC would be too small resulting

in an underpowered study. The only case in which it would not make a difference if the researcher borrowed from reading or mathematics is if the estimates are the same. This was only the case for Massachusetts, grade 5.

Table 3.1

Empirically Estimated ICCs from Two-Level Models with Students Nested in Schools

Source[a]	Mathematics				Reading		
	5	8	10	11	5	8	10
National Educational Longitudinal Study, 1988 (Hedges & Hedberg, 2007)	0.216	0.185	0.234	0.138	0.263	0.197	0.183
District A (Bloom et al., 2007)	0.20	0.16	0.13		0.25	0.18	0.15
District B (Bloom et al., 2007)	0.19				0.15		
District C (Bloom et al., 2007)	0.17	0.27	0.25		0.20	0.23	0.29
District E (Bloom et al., 2007)	0.18				0.12		
Longitudinal Evaluation of School Change and Performance (Schochet, 2008)	0.18				0.21		
21st Century Community Learning Centers Program (Schochet, 2008)	0.17				0.09		
Massachusetts Department of Education (Hedberg & Hedges, 2011)	0.239	0.276			0.239	0.249	
Florida Elementary Schools Data (Zhu et al., 2012)	0.132				0.109		
North Carolina Elementary Schools Data (Zhu et al., 2012)	0.118				0.090		
Hawaii State Assessment Total Reading[b] (Brandon et al., 2013)					0.177	0.137	0.136

[a] Decimal length of reported ICC values varies by source, and is preserved in this summary table.
[b] Hawaii State Assessment Total Reading ICCs are a 95% confidence interval upper-bound.

It is also important to note that the ICCs within subject vary across the studies. Looking down the column for grade 5, reading ICCs range 0.09-0.26, while ICCs in mathematics range 0.12-0.24. The data source for the studies differs which may explain some of this variability. The data source for the studies include: a collection of national samples (Hedges & Hedberg, 2007), all districts in a single state (Brandon, Harrison, & Lawton, 2013; Hedberg & Hedges, 2011; Zhu et al., 2012), single districts in multiple states (Bloom et al., 2007), and multiple districts in multiple states (Schochet, 2008). Hence it is critical that a researcher carefully considers the most relevant data source in selecting the appropriate ICC.

Challenges with Borrowing R^2 Values

The importance of the use of covariate sets to increase the precision of a study has been well established. The covariate set that rises to the top for mathematics and reading outcomes in terms of the explanatory power is the one-year lagged same subject, student-level pretest (Bloom et al., 2007; Hedges & Hedberg, 2007; Zhu et al., 2012). The key challenge for science studies is that the one-year lagged same subject, student-level pretest often does not exist. Assuming the testing pattern of grades 5, 8, 10, and 11, a one-year lagged same subject, student-level pretest would only be available for grade 11. For grade 10, a two-year lag would be the closest available student-level pretest, for grade 8, a three-year lag, and there is no available student-level pretest for grade 5. Bloom et al. (2007) show that for mathematics and reading, the explanatory power of student-level pretests reduces slightly as the number of years between pretest and posttest

increases. Hence it is critical to examine the specific covariate sets available for science to determine which are the most powerful.

Other covariate sets besides the one-year lagged same subject student-level pretest have also been explored in reading and mathematics. Researchers have found that the explanatory power of school-level pretests can be nearly as effective as student-level pretests (Bloom et al., 2007; Gargani & Cook, 2005; Jacob, Goddard, & Kim, 2014). Also, the use of a cross-subject pretest covariate (i.e., reading achievement pretest with a mathematics achievement outcome) was found to be helpful. For science, school-level science pretests are available for all grades and thus may offer a powerful alternative for grades in which there is no one-year lagged student-level science pretest. Cross-subject pretests may also be important for science studies since a one-year lagged student-level mathematics and reading pretest is available for all grades.

Research Questions

As noted above, no systematic investigations of science ICCs exists. Hence the first question we ask is:

1. What are unconditional ICCs for science achievement outcomes?

Given the practice of borrowing ICCs from reading and mathematics, we are also interested in how science ICCs compare to mathematics and reading ICCs. As we saw in Table 3.1, ICCs can vary greatly across data sources. Hence we use the same data set to calculate reading and mathematics ICCs in order to address our second research question:

2. How do the empirical estimates of ICCs for science achievement compare to those for reading and mathematics achievement?

Finally, we know the importance of the use of covariate sets to increase the power of a study. The unique testing patterns for science do not allow for a one-year lagged same subject, student-level pretest for all grades. However, one-year lagged school-level pretests and one-year cross subject student-level pretest are available for all grades. Hence our third question is:

3. For the grade levels in which science is tested, which covariate sets explain the most variance?

The remainder of this paper is organized as follows. First, we provide a description of the data used to address the three focal research questions. Then we describe the two models we use to estimate the parameters: the two-level hierarchical linear model (HLM) with students nested within schools and the three-level HLM with students nested within schools nested within districts. In the results section, we present our estimates for the unconditional ICCs for science, reading, and mathematics followed by the explanatory power of the covariate sets specific to science. Next, we collectively consider the various covariate models and demonstrate the use of the empirical estimates using a brief example. Finally, we present our conclusions and address the limitations of this study.

Method

Data

Data from the Texas Education Agency (TEA) for the State of Texas was obtained for 5 academic years beginning in the 2006-07 academic year. The data

included student-level achievement data for science, mathematics, and reading from the Texas Assessment of Knowledge and Skills (TAKS), student demographic information (e.g., gender, race, socio-economic status), and school and district identifiers. In Texas, as was noted above, science is tested only in grades 5, 8, 10, and 11, while mathematics and reading are tested in all grades 3-11. Testing for science using the TAKS occurs annually in April.

In accordance with the Family Educational Rights and Privacy Act (FERPA), the TEA masked data according to the lowest level of clustering. When fewer than five individuals exist in any single group within a cluster (school), achievement scores for every student in that group are masked. Masking occurs less often in higher grades because school size generally grows as the grade increases.

Table 3.2 describes the number of students, schools, and districts by grade for 2007-2011 for the unconditional analysis. Students were removed from the analysis due to the masking process as well as the data cleaning process. The masking makes it is difficult to account for certain demographic or other testing information (e.g., special education students and students receiving a test accommodation) because the act of identifying this information produces a disproportionately large incidence of masking within these subgroups. However, these students are often excluded from research studies. Therefore, we have removed students with these identifiers from the analysis.[7] In Table 3.2, we report incidence of masking as a percentage of the remaining data after

[7] Students excluded from the analysis solely because of a testing accommodation represent approximately 4% of the non-masked, cleaned data. A sensitivity analysis comparing the results with and without the removal of these students shows very little difference.

cleaning. The incidence of masked data in the cleaned Texas dataset is approximately 16%; however, this varies by grade.

Table 3.2

Science Achievement Unconditional Model Sample Sizes for Student, School, and District

Grade	Year	% of Data Removed by Cleaning	% of Cleaned Data that is Masked	Total Students	Total Schools	Total Districts
Grade 5	2007	16.4	22.5	215,443	3,414	932
	2008	20.2	22.0	216,480	3,497	947
	2009	20.1	21.7	221,713	3,581	955
	2010	19.0	28.1	210,770	3,625	954
	2011	20.0	26.7	218,712	3,681	952
Grade 8	2007	15.8	13.2	247,800	1,601	934
	2008	16.8	12.9	241,312	1,610	926
	2009	16.5	12.4	252,122	1,646	943
	2010	15.3	17.4	242,141	1,664	937
	2011	15.9	16.2	249,608	1,704	942
Grade 10	2007	16.8	11.5	235,828	1,279	931
	2008	17.2	11.7	236,411	1,312	944
	2009	16.3	11.3	241,548	1,324	941
	2010	15.7	15.9	231,799	1,329	923
	2011	16.8	14.3	239,716	1,348	925
Grade 11	2007	15.9	11.4	206,076	1,229	894
	2008	15.6	11.5	208,188	1,254	908
	2009	14.8	11.2	218,876	1,307	930
	2010	14.2	15.4	216,260	1,321	917
	2011	15.9	14.3	217,526	1,326	899

Data from the TEA were cleaned to obtain a consistent set of usable data across years. The analysis was performed on non-masked, non-special education students with a valid, unique student identification number that took the English version of the TAKS in the standard administration setting that were scored. Students with non-valid or duplicate student identification numbers, as well as students taking an alternate version of the TAKS (i.e., TAKS [Accommodated], TAKS-Modified, and TAKS-Alternative) or otherwise requiring an accommodation (i.e., presentation, response, setting, timing and scheduling, and oral administration) were removed from the sample.

In all years, indicator variables were generated from a set of demographic variables including gender, race, socio-economic status (SES), and limited English proficiency (LEP) status. Since the race variable included five or more categories, four indicator variables were created to denote each of five racial identifiers. In 2007-2010, race was collected with five sub-categories (American Indian or Alaskan Native, Asian or Pacific Islander, African American, Hispanic, and White, not of Hispanic Origin). In 2011, race was collected with seven sub-categories (Hispanic/Latino, American Indian or Alaskan Native, Asian, Black or African American, Native Hawaiian or Other Pacific Islander, White, or Two or More Races). The 2011 coding was collapsed to fit with the 2007-2010 coding structure by merging Asian and Native Hawaiian or Other Pacific Islander, and eliminating the records coded as two or more races. A proxy for SES was developed based on free or reduced-priced lunch status. The SES variable was coded 1 for any student eligible for free or reduced-priced meals or other economic disadvantage, and coded 0 when the student was not identified as economically disadvantaged. An LEP

variable was coded 1 for any student that had ever been identified as LEP, and coded 0 otherwise. Collectively, the indicator variables for gender, race, SES, and LEP make up the set of demographic covariates.

HLM Models

There are two primary models of interest for this study. The first is the two-level HLM with students nested within schools; this model treats school as a random effect, but ignores the district-level. The second is the three-level HLM with students nested within schools nested within districts. In the three-level HLM, both school and district are included as random effects. We present the theoretical framework for both the two-level and three-level HLM.

The unconditional two-level HLM. 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:

$$Y_{ij} = \beta_{0j} + r_{ij} \qquad r_{ij} \sim N(0, \sigma^2), \qquad \textbf{[21]}$$

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

$$\beta_{0j} = \gamma_{00} + u_{0j} \qquad u_{0j} \sim N(0, \tau_{00}), \qquad \textbf{[22]}$$

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,

$$\rho = \frac{\tau_{00}}{\tau_{00} + \sigma^2}. \quad \textbf{[23]}$$

The standard error estimate of the ICC, assuming large sample sizes, is given by Hedges, Hedberg, and Kuyper (2012) and Donner and Koval (1982),

$$SE(\rho) = \sqrt{\frac{(1-\rho)^2 v_2}{(\sigma^2 + \tau_{00})^2}}, \quad \textbf{[24]}$$

where v_2 is the variance of the variance component estimate of τ_{00}. For large sample sizes Equation [4] is asymptotically equivalent to similar formulas given by Fischer (1925) as well as Donner and Koval (1980).

The unconditional three-level HLM. The unconditional model for the three-level HLM with students (Level 1) nested within schools (Level 2) nested within districts (Level 3) is as follows. The Level 1 or student-level model is:

$$Y_{ijk} = \pi_{0jk} + e_{ijk} \qquad e_{ijk} \sim N(0, \sigma^2) \quad \textbf{[25]}$$

where Y_{ijk} is the outcome for individual $i \in \{1, \ldots, n_{jk}\}$, in school $j \in \{1, \ldots, J_k\}$, in district $k \in \{1, \ldots, K\}$, π_{0jk} is the mean achievement at school *j* in district *k*, and e_{ijk} is a random student effect, which is assumed to be normally distributed with a mean of 0 and homogenous variance σ^2. Therefore, σ^2 is the variance in achievement among students within schools. The Level 2 or school-level model is:

$$\pi_{0jk} = \beta_{00k} + r_{0jk} \qquad r_{0jk} \sim N(0, \tau_\pi), \qquad \textbf{[26]}$$

where β_{00k} is the mean in district *k*, and r_{0jk} is the random school effect, which is assumed to be normally distributed with a mean of 0 and homogenous variance τ_π. Therefore, τ_π is the variance in mean achievement among schools within districts. The Level 3 or district-level model is:

$$\beta_{00k} = \gamma_{000} + u_{00k} \qquad u_{00k} \sim N\left(0, \tau_\beta\right), \qquad \textbf{[27]}$$

where γ_{000} is the grand mean, and u_{00k} is a random district effect, which is assumed to be normally distributed with a mean of 0 and homogenous variance τ_β. Therefore τ_β is the variance in mean achievement among districts.

In the three-level HLM, there are two ICCs. The school-level ICC, or proportion of total variance that exists among schools within districts is

$$\rho_2 = \frac{\tau_\pi}{\tau_\beta + \tau_\pi + \sigma^2}. \qquad \textbf{[28]}$$

The district-level ICC, or proportion of total variance that exists among districts is

$$\rho_3 = \frac{\tau_\beta}{\tau_\beta + \tau_\pi + \sigma^2}. \qquad \textbf{[29]}$$

The respective standard errors of ρ_2 and ρ_3 in the large sample, balanced (i.e., $\forall k, J_k = J$) three-level model are:

$$SE(\rho_2) = \sqrt{\frac{[J(1-\rho_2)^2 + 2\rho_2(1-\rho_2)]v_2 + J\rho_2^2 v_3}{J\left(\tau_\beta + \tau_\pi + \sigma^2\right)^2}}, \qquad \textbf{[30]}$$

and

$$SE(\rho_3) = \sqrt{\frac{[J\rho_3^2 + 2\rho_3(1-\rho_3)]v_2 + J(1-\rho_3)^2 v_3}{J\left(\tau_\beta + \tau_\pi + \sigma^2\right)^2}}, \quad \textbf{[31]}$$

where v_2 and v_3 are variances of the variance component estimates of τ_π and τ_β, respectively, and J is assumed to be the harmonic mean number of schools per district (Hedges et al., 2012).

Proportion of variance explained by the covariate sets. As unconditional models are modified to include individual-level and cluster-level covariates, variance is explained. In this study, Level 1 covariates are student-level data (e.g., gender, scores on prior test), Level 2 covariates represent averages across students within schools, and Level 3 covariates represent averages across schools within districts. Because of the hierarchical structure, Level 1 covariates can theoretically be used without a corresponding Level 2 covariate; however, if a Level 1 covariate is used in our model, we aggregate it to be used as a Level 2 covariate as well, and if a Level 2 covariate is used in a model, we also aggregate it to be used at Level 3. Below is the model for a two-level HLM with covariates at Level 1 and Level 2. The model for a three-level HLM is excluded since this model can be easily extended from the two-level HLM case.

In the conditional two-level HLM, the new Level 1, or student-level model is:

$$Y_{ij} = \beta_{0j} + \sum_q \beta_{qj} X_{qij} + r_{ij} \qquad r_{ij} \sim N\left(0, \sigma^2_{|X_Q}\right) \quad \textbf{[32]}$$

for $i \in \{1, 2, \ldots, n_j\}$ students per school and $j \in \{1, 2, \ldots, J\}$ schools, where Y_{ij} is the outcome for student *I* in school *j*, β_{0j} is the mean for school *j*, X_{qij} is the value of the q^{th} student-level covariate $q \in \{1,2, \ldots, Q\}$ for student *I* in school *j*, β_{qj} is the student-level

coefficient associated with the q^{th} student-level covariate for school j, r_{ij} is a random student effect or the residual error associated with each student, conditional on the Q covariates, which is assumed to be normally distributed with mean 0 and homogeneous variance $\sigma^2_{|\boldsymbol{X}_Q}$. Therefore, $\sigma^2_{|\boldsymbol{X}_Q}$ is the residual variance in achievement among students within schools after adjusting for the Q Level 1 covariates. The new Level 2 or school-level model is:

$$\beta_{0j} = \gamma_{00} + \sum_s \gamma_{0s} W_{sj} + u_{0j} \qquad u_{0j} \sim N\left(0, \tau_{|\boldsymbol{W}_S}\right)$$
$$\beta_{qj} = \gamma_{q0}, \forall q \in \{1,2,\dots,Q\} \qquad \textbf{[33]}$$

where γ_{00} is the grand mean of the adjusted outcome measure, W_{sj} is the value of the s^{th} school-level covariate $s \in \{1,2,\dots,S\}$ for school j, γ_{0s} is the fixed school-level coefficient associated with the s^{th} school-level covariate, u_{0j} is the residual error associated with each school, conditional on the S covariates, and $\tau_{|\boldsymbol{W}_s}$ is the residual variance among schools after adjusting for the S Level 2 covariates. Slope values for the Q Level 1 covariates are assumed to have fixed effects across schools, and γ_{q0} represents the fixed student-level coefficient associated with the q^{th} student-level covariate.

The proportion of variance reduced at each level by the inclusion of the covariate sets, denoted R^2, is calculated using the results from the unconditional model and conditional models containing one or more covariates. The R^2 at Level 1 and Level 2 are calculated as follows:

$$R^2_{L1} = \frac{\sigma^2 - \sigma^2_{|\boldsymbol{X}_Q}}{\sigma^2}, \qquad \textbf{[34]}$$

and

$$R^2_{L2} = \frac{\tau_{00} - \tau_{|\boldsymbol{W}_S}}{\tau_{00}}. \quad \textbf{[35]}$$

Data Analysis

To empirically estimate ICC values from the TEA dataset we use a modified Stata (StataCorp, 2011) program which uses the XTMIXED function with a restricted maximum-likelihood estimation procedure for the estimates and the ICCVAR function assuming a balanced design to compute bounds on the ICC values (Hedges et al., 2012; Hedberg, 2012). We estimate the unconditional ICCs for science, reading, and mathematics achievement for the same set of students in order to compare values across the subject areas.

We use the same program to calculate R^2 values. We strategically examined the following covariate sets: (a) demographics, (b) the most recent student-level science pretest, (c) the one-year lagged student-level reading pretest, (d) the one-year lagged student-level mathematics pretest, (e) the one-year lagged school-level science pretest, (f) the one-year lagged school-level reading pretest, and (g) the one-year lagged school-level mathematics pretest. The most recent student-level science pretest may be a one-, two-, or three-year lag depending on the grade. Student-level reading and mathematics pretest covariates are always available with a one-year lag as these subjects are tested in every grade. School-level covariates, regardless of subject, are always available with a one-year lag, and demographic covariates can be added to any model. The full list of relevant covariates is presented in Table 3.3.

Table 3.3

Covariate Definitions

Models	Covariate Definitions
y_d	Demographics-only
$y_{s\text{-}t}$	Same student scores in science lagged *t*-year(s), $t \in \{1, 2, 3\}$
$y_{m\text{-}1}$	Same student scores in mathematics lagged one year
$y_{r\text{-}1}$	Same student scores in reading lagged one year
$Y_{s\text{-}1}$	Mean school scores in science for the same grade lagged one year
$Y_{m\text{-}1}$	Mean school scores in mathematics for the same grade lagged one year
$Y_{r\text{-}1}$	Mean school scores in reading for the same grade lagged one year

Note. All pretest models can be run with or without the set of demographic variables.

There are of course other possibilities, such as the two-year lagged student-level reading or mathematics pretest or the two-year lagged school-level science pretest, but we maintain these are not likely to be used. As demonstrated by Bloom et al. (2007) the explanatory power decreases for longer length lags. Since the one-year lagged school-level science pretest is available, we do not consider the two-year (or more) lagged school-level science pretest. Likewise, the two-year (or more) lagged student-level reading and mathematics pretests are unnecessary because in Texas students are tested in these subjects annually, which ensures a one-year lagged student-level pretest is available for both reading and mathematics.

Results

The results are organized as follows. We begin by presenting the unconditional ICCs for the two-level and three-level HLMs for each grade and subject. Next, we

examine the percentage of variance in science achievement explained with particular pretest covariates and with demographic characteristics for each grade. In all cases, the presentation of a single ICC or R^2 value represents an average across the years in which the statistic can be calculated.

Unconditional Model

Unconditional ICC and standard error estimates for science, reading, and mathematics achievement are presented in Table 3.4 for both the two-level and three-level HLM. The average unconditional ICCs in the two-level HLM range from 0.172 to 0.196 for science. For the three-level HLM, the school-level science ICCs range from 0.104 to 0.136 and the district-level ICC ranges from 0.055 to 0.079, depending on grade. This suggests that approximately one-third of the variance at the school-level actually occurs at the district-level. While it is unlikely that a researcher would design a CRT that involved random assignment at the district-level, an estimate of the school-level and district-level ICCs do provide an approximate bound on the amount of variance that could exist at the school-level if a within-district design is utilized when planning a CRT, which is a more common approach.

Table 3.4

Average Unconditional ICCs for a Two-Level and Three-Level HLM by Grade for Science, Reading, and Mathematics Achievement, 2007-2011

Achievement Outcome Subject	Grade	Two-Level HLM		Three-Level HLM			
		ICC	*SE*	ICC_{L2}	*SE*	ICC_{L3}	*SE*
Science	5	0.191	0.004	0.118	0.003	0.079	0.007
	8	0.172	0.005	0.104	0.005	0.060	0.007
	10	0.196	0.007	0.136	0.008	0.055	0.008
	11	0.191	0.007	0.127	0.008	0.059	0.008
Reading	5	0.156	0.004	0.097	0.003	0.050	0.005
	8	0.099	0.004	0.060	0.003	0.031	0.004
	10	0.140	0.006	0.100	0.007	0.037	0.007
	11	0.122	0.006	0.092	0.006	0.025	0.005
Mathematics	5	0.168	0.004	0.105	0.003	0.067	0.006
	8	0.163	0.005	0.103	0.005	0.053	0.006
	10	0.169	0.006	0.124	0.007	0.042	0.007
	11	0.172	0.007	0.119	0.007	0.049	0.007

Unconditional ICCs can also be compared across subjects. In the two-level HLM, the average unconditional ICCs for science are larger than the unconditional ICCs for reading for all grades. The most dramatic differences occur in the middle and high school grades. The science ICC in grade 8 is 0.172 and the corresponding reading ICC is 0.099. In essence, the science ICC is almost twice as large as the reading ICC. Compared to the mathematics ICC, the science ICC tends to be slightly larger, although the margin is much smaller than that between reading and science. In the three-level HLM, a similar pattern exists for the average unconditional school-level ICCs as well as the

unconditional district-level ICCs with science ICCs consistently being larger than reading ICCs and the same as or slightly larger than mathematics ICCs.

In all cases, ICC values appear to decrease somewhat between grade 5 and 8 and then increase again for the high school grades, though considering the standard errors for each grade negates or nearly negates this trend for mathematics and science. We hypothesize that one reason for a drop between grade 5 and 8 may be the vast difference in the number of elementary schools versus middle schools. Consolidation of students at the school-level generally translates into more heterogeneity among students within schools, and hence more homogeneity among schools. Since there are over twice as many elementary schools as middle schools, we would expect larger ICCs in grade 5 than in grade 8. This is consistent to a finding from Hedges and Hedberg (2007) that ICCs generally decrease slightly as grade increase for both mathematics and reading.

The subsequent increase in average ICCs between grade 8 and grades 10 and 11 is more puzzling, though not unique to science; a somewhat similar pattern can be found in Hedges and Hedberg (2007) in that middle school unconditional ICCs for mathematics and reading are often smaller than elementary and high school ICCs. However, without more examples of science ICCs, for example from other grades and states, it is difficult to identify the true source(s) of the larger ICCs in these grades.

Models with Covariate Sets

As discussed earlier, the unique testing patterns in science in Texas means that not all covariate sets are available for each grade. Table 3.5 presents the grades in which each covariate set can be run for each year of data. For example, for the earliest year,

2007, no pretest covariates are available in the data, and the only covariate set available is demographics. In 2008, the set of demographics is available for each grade as well as the cross-subject tests lagged one-year at the student-level and the school-level tests in all subjects lagged one year. In 2008, a one-year lagged student-level science pretest is also available for grade 11. Note that in subsequent years, a one-year lagged student-level science pretest is available for grade 11, a two-year lagged student-level science pretest is available for grade 10 and a three-year lagged student-level science pretest is available for grades 8 and 11. In our findings, we present the R^2 values as averages across years in which the covariate sets are run.

Using science achievement as the outcome, we estimate R^2 values for the demographics-only model and a total of six pretest models. Note that the one year lagged-student level pretest is considered one model, the most recent science pretest, regardless of the number of lagged years. We also examined each of the pretest models with demographics. Hedges and Hedberg (2007) and Bloom et al. (2007) show results from models with only demographics, and with only a pretest can explain a considerable amount of variance, but that models with both pretests and demographics produce little value beyond when only a pretest is used. Our findings echo this result, especially for models with one-year lagged student-level pretest covariates. Because including the demographics covariates can use up valuable degrees-of-freedom, and does not considerably affect our conclusions, in the interest of consistency and space, we limit our presentation and interpretation to the demographics-only model and the six models with

only a pretest. Results of the models with a pretest and demographics are not elaborated on, but for completeness can be found in the Appendix.

Table 3.5

Grades in Which Data are Available Across Years for Relevant Models

Models[a]	Outcome Years				
	2007	2008	2009	2010	2011
y_d	5,8,10,11	5,8,10,11	5,8,10,11	5,8,10,11	5,8,10,11
y_{s-1}		11	11	11	11
y_{s-2}			10	10	10
y_{s-3}				8, 11[b]	8, 11[b]
y_{m-1}		5,8,10,11	5,8,10,11	5,8,10,11	5,8,10,11
y_{r-1}		5,8,10,11	5,8,10,11	5,8,10,11	5,8,10,11
Y_{s-1}		5,8,10,11	5,8,10,11	5,8,10,11	5,8,10,11
Y_{m-1}		5,8,10,11	5,8,10,11	5,8,10,11	5,8,10,11
Y_{r-1}		5,8,10,11	5,8,10,11	5,8,10,11	5,8,10,11

[a] Relevant models are defined by their covariates: y_d is the demographics-only model; y_{s-1} is the same student scores in science lagged one year; y_{s-2} is the same student scores in science lagged two years; y_{s-3} is the same student scores in science lagged three years; y_{m-1} is the same student scores in mathematics lagged one year; y_{r-1} is the same student scores in reading lagged one year; Y_{s-1} is the mean school scores in science for the same grade lagged one year; Y_{m-1} is the mean school scores in mathematics for the same grade lagged one year; Y_{r-1} is the mean school scores in reading for the same grade lagged one year. All models with a pretest covariate can be run with or without the set of demographic variables.
[b] A three-year lagged student-level pretest in science can be computed for both grades 8 and 11, but we do not present the calculation for grade 11 in our tables of findings because a one-year lagged student-level pretest in science is available for that grade, and would be preferable.

Demographics only. Table 3.6 shows the effect of including only the demographics covariate set. The demographics covariates alone account for a considerable amount of variance in both the two-level and three-level HLMs of science achievement. In the two-level HLM, adding the demographics covariate set explains

52.7-61.5% of the school-level variance depending on the grade. In the three-level HLM, the demographics covariate set explains 50.7-66.3% of the school-level variance, and 45.8-66.5% of the district-level variance.

Table 3.6

Average R^2 for a Two-Level and Three-Level HLM of Science Achievement with Demographics Covariates by Grade, 2007-2011

	Two-Level HLM		Three-Level HLM		
Grade	R^2_{L1}	R^2_{L2}	R^2_{L1}	R^2_{L2}	R^2_{L3}
5	0.103	0.527	0.103	0.507	0.498
8	0.134	0.615	0.134	0.663	0.458
10	0.128	0.615	0.128	0.609	0.665
11	0.130	0.598	0.130	0.616	0.618

Most recent student-level pretests. We consider the effect of a student-level pretest using the most recent student-level pretest in science, reading, and mathematics. Recall that student-level science pretests are not available for grade 5. For grade 8, student scores in grade 5 are the most recent science pretest, a three-year lag. In grade 10, student scores in grade 8 are the most recent science pretest, a two-year lag. In grade 11, student scores in grade 10 are the most recent science pretest, a one-year lag. For all grades, a mathematics or reading pretest is available as a one-year lag.

Table 3.7 shows the results for the student-level pretests. The results suggest that the researcher's ability to explain variance by adding a student-level science pretest is strong, but diminishes as the lag increases on the pretest. In the two-level HLM, a one-year lagged student-level science pretest (grade 11) explains 91.2% of the school-level

variance, whereas the science pretest with a two-year (grade 10) lag or three-year (grade 8) lag explains 80.6 and 64.0% of the school-level variance, respectively. The researcher's ability to explain student-level variance diminishes proportionally as well. For the three-level HLM, 92.2% of the district-level variance, 90.3% of the school-level variance, and 50.5% of the student-level variance is explained using a one-year lagged (grade 11) student-level science pretest, but as the lag length increases, the percentage of variance explained decreases considerably.

Table 3.7

Average R^2 for a Two-Level and Three-Level HLM of Science Achievement with the Most Recent Student-Level Science, Reading, or Mathematics Pretest Covariate by Grade, 2008-2011

Pretest Subject	Grade	Two-Level HLM		Three-Level HLM		
		R^2_{L1}	R^2_{L2}	R^2_{L1}	R^2_{L2}	R^2_{L3}
Science	5	-	-	-	-	-
	8	0.297	0.640	0.297	0.630	0.565
	10	0.470	0.806	0.470	0.817	0.740
	11	0.505	0.912	0.505	0.903	0.922
Reading	5	0.268	0.634	0.268	0.558	0.713
	8	0.319	0.745	0.319	0.758	0.605
	10	0.167	0.664	0.167	0.688	0.550
	11	0.191	0.707	0.191	0.678	0.729
Mathematics	5	0.270	0.628	0.270	0.522	0.747
	8	0.413	0.754	0.413	0.714	0.741
	10	0.439	0.839	0.439	0.865	0.718
	11	0.445	0.817	0.445	0.827	0.758

The result that less variance is explained with a longer lag in the student-level science pretest is confounded with the fact that grade level and number of lag years are perfectly aligned, and so it is unclear whether a grade-effect or a lag-effect is associated with this result. In an effort to clarify this finding further, we estimated the R^2 value using a three-year lagged student-level science pretest covariate for grade 11[8] and compared it to the corresponding statistic for grade 8. For the two-level HLM, the pretest explains more school-level variance for grade 11 (76.9%) than for grade 8 (64.0%), and a similar pattern exists for school-level variances the three-level HLM. The difference in variance explained across grades with the same length lag suggests to some extent that a grade effect may also be present.

We can also compare the results for the most recent student-level science pretest to the one-year lagged student-level cross subject pretest. In the two-level HLM, when the one-year lagged student-level science pretest is available (grade 11), the proportion of school-level variance explained, 91.2%, is greater than the one-year lagged student-level reading or mathematics pretest, 70.7 and 81.7%, respectively. However, in grade 10 when only a two-year lagged science pretest is available, the one-year lagged mathematics pretest explains slightly more variance than the two-year lagged science pretest, 83.9 and 80.6%, respectively. However, the one-year lagged reading pretest is not as powerful and only explains 66.4% of the variance at grade 10. In grade 8, when only a three-year lagged science pretest is available, one-year lagged student-level pretests in both reading

---

[8] The three-year lagged student-level science pretest for grade 11 is possible in the Texas dataset (see Table 3.5), but ultimately an unnecessary model given that a one-year lagged student-level pretest is available for that grade.

and mathematics explain more variance than the science pretest, 74.5, 75.4, and 64.0%, respectively. In the three-level HLM, a similar pattern exists for both the district and school level. For grades 11 and 10, the most recent student-level science pretest explains more variance than the one-year lagged student-level reading or mathematics pretest at the district level. At the school level, the one-year lagged student-level mathematics pretest explains slightly more variance than the two-year lagged science pretest, 86.5 and 81.7%, respectively. When there is a three-year lagged student-level science pretest, grade 8, the one-year lagged reading or mathematics pretest explain more variance than the science pretest.

One-year lagged school-level pretests. Often student-level pretest scores are too expensive or otherwise not possible to obtain, but school-level pretest covariates are readily available. The extent that a one-year lagged school-level science, reading, or mathematics pretest covariate explains variance in science achievement is presented in Table 3.8. For one-year lagged school-level pretests, we report only the covariates' contribution to explaining school-level and, when applicable, district-level variances. Regardless of whether the model is a two-level or three-level HLM, student-level variance can only be explained by student-level covariates and thus in this case, the level-one variance remains unchanged.

For the two-level HLM, a one-year lagged school-level science pretest explains 67.5-86.8% of the school-level variance in science achievement depending on grade. For the three-level HLM, 54.6-86.5% of the school-level variance, and 83.6-91.7% of the district-level variance in science achievement is explained, depending on grade.

Table 3.8

Average R^2 for a Two-Level and Three-Level HLM of Science Achievement with a One-Year School-Level Science, Reading, or Mathematics Pretest Covariate by Grade, 2008-2011

Pretest Subject	Grade	Two-Level HLM	Three-Level HLM	
		R^2_{L2}	R^2_{L2}	R^2_{L3}
Science	5	0.675	0.546	0.917
	8	0.802	0.739	0.856
	10	0.868	0.858	0.859
	11	0.866	0.865	0.836
Reading	5	0.582	0.472	0.713
	8	0.658	0.619	0.623
	10	0.629	0.634	0.525
	11	0.584	0.583	0.586
Mathematics	5	0.569	0.440	0.755
	8	0.671	0.630	0.617
	10	0.761	0.797	0.581
	11	0.783	0.803	0.679

Conceptually, a school-level science pretest is preferable to a school-level cross-subject pretest because it is theoretically more justifiable. Additionally, based on the work of Bloom et al. (2007), empirically we know that same-subject pretests tend to have more explanatory power than cross-subject pretests, at least for mathematics and reading outcomes. Thus, our inclusion of cross-subject school-level pretest models may seem odd at first, given that science is tested annually in each of the specified grades, and therefore always available. However, we note there is always a chance science testing was otherwise not conducted for the same grade in the prior year (e.g., change in testing

patterns, lack of funding) or the results are not readily available, and so in these instances a school-level reading or mathematics pretest could be a logical choice.

As one would expect, across all grades, a school-level science pretest explains more variance than either mathematics or reading. The explanatory power of school-level mathematics and reading pretests was similar for grade 5 and 8. However, in grades 10 and 11, the explanatory power of the school-level mathematics pretest exceeds the power of the school-level reading pretest by somewhat larger margins. For example, in the two-level model, the reading pretest explains 58.4% of the variance in science achievement whereas the mathematics pretest explains 78.3% of the variance in science achievement.

All covariate options. Thus far, we examined demographics, student-level pretest covariate sets, and school-level pretest covariate sets. In Table 3.9, we summarize all of the findings. Note that the covariate set explaining the most variance at the highest level is presented in bold and varies by grade. For grade 11, where a one-year student-level science pretest is available, this is the most powerful covariate set for both the two-level and three-level HLM. In grades 8 and 10, where a student-level pretest is lagged more than one year, the most variance is explained by a school-level science pretest for both models. In grade 5, where no student-level science pretest is available, the school-level science pretest explains the most variance for the two-level and three-level HLM.

Table 3.9

Maximum Science Achievement R^2 in the Highest Level of Nesting for Two-Level HLM and Three-Level HLM with Relevant Covariate Sets by Grade

Type of HLM	Models[a, b]	Highest Level of Nesting	R^2			
			Grade 5	Grade 8	Grade 10	Grade 11
Two-Level	y_d	L2	0.527	0.615	0.615	0.598
	$y_{s\text{-}t}$	L2	-	0.640	0.806	**0.912**
	$y_{m\text{-}1}$	L2	0.628	0.754	0.839	0.817
	$y_{r\text{-}1}$	L2	0.634	0.745	0.664	0.707
	$Y_{s\text{-}1}$	L2	**0.675**	**0.802**	**0.868**	0.866
	$Y_{m\text{-}1}$	L2	0.569	0.671	0.761	0.783
	$Y_{r\text{-}1}$	L2	0.582	0.658	0.629	0.584
Three-Level	y_d	L3	0.498	0.458	0.665	0.618
	$y_{s\text{-}t}$	L3	-	0.565	0.740	**0.922**
	$y_{m\text{-}1}$	L3	0.747	0.741	0.718	0.758
	$y_{r\text{-}1}$	L3	0.713	0.605	0.550	0.729
	$Y_{s\text{-}1}$	L3	**0.917**	**0.856**	**0.859**	0.836
	$Y_{m\text{-}1}$	L3	0.755	0.617	0.581	0.679
	$Y_{r\text{-}1}$	L3	0.713	0.623	0.525	0.586

[a] Relevant models are defined by their covariates: y_d is the demographics-only model; $y_{s\text{-}1}$ is the same student scores in science lagged one year; $y_{s\text{-}2}$ is the same student scores in science lagged two years; $y_{s\text{-}3}$ is the same student scores in science lagged three years; $y_{m\text{-}1}$ is the same student scores in mathematics lagged one year; $y_{r\text{-}1}$ is the same student scores in reading lagged one year; $Y_{s\text{-}1}$ is the mean school scores in science for the same grade lagged one year; $Y_{m\text{-}1}$ is the mean school scores in mathematics for the same grade lagged one year; $Y_{r\text{-}1}$ is the mean school scores in reading for the same grade lagged one year. All models with a pretest covariate can be run with or without the set of demographic variables; results shown here for pretest models reflect only pretest models without demographics variables.

[b] For student-level science pretests, $t \in \{1, 2, 3\}$. In grades 11, 10, and 8, respectively, t=1, t=2, and t=3. Maximum variance explained in each grade excluding covariate models with both a pretest and demographics are highlighted in bold.

Application

We now present an example to illustrate how our findings can be utilized. Numerous examples of how to appropriately power CRTs exist in the literature (Bloom et al., 2007; Hedges & Hedberg, 2007); we build on this foundation to illustrate that the choice of pretest is contextual, and can lead to considerable differences with regard to optimally designing a CRT.

Suppose that a team of researchers are designing a CRT to test the effectiveness of a science intervention aimed at 8^{th} graders. They propose a two-level CRT, with students nested within schools. They want to design a study that is powered at 0.80 to detect an effect of 0.20, with a significance level of 0.05. They plan to select 100 students from each school. Based on the findings presented in this paper, they assume the unconditional ICC is 0.172. The researchers are unsure which covariate set will be the most powerful and hence result in the smallest number of schools needed to adequately power the study. Because the most recent science test for students was three years prior, the researchers consider the three-year lagged student-level science pretest. However, they also have the one-year lagged student-level mathematics pretest which they think may be better since it is only lagged one-year. Finally, they consider the one-year lagged school-level science pretest.

Using the Optimal Design Plus program (Spybrook, Bloom, Condon, Martinez, & Raudenbush, 2011), the researcher computes the number of schools needed for a design with no covariates, the two-year lagged student-level science pretest, the one-year lagged student-level mathematics pretest, and the one-year lagged school-level science pretest.

Without a covariate, approximately 146 schools are required to power the study. According to Table 3.7, the three-year lagged student-level science pretest accounts for 29.7% of the variance at level-one and 64.0% of the variance at level-two. Using these estimates, approximately 56 schools total are necessary. The second option, the one-year lagged student-level mathematics pretest accounts for 41.3% of the variance at level-one and 75.4% of the variance at level-two (see Table 3.7). Under these assumptions, a total of approximately 40 schools are needed. The one-year lagged school-level science pretest is a third option available to researchers. According to Table 3.8, 80.2% of the school-level variance in science achievement is explained by this covariate. In this case, the total number of schools is approximately 36. In this particular example, the one-year student-level mathematics pretest was better than the three-year lagged student-level science pretest. However, the school-level pretest was more powerful than either of the two student-level pretest options and hence yielded the smallest number of schools to achieve power of 0.80 to detect an effect of 0.20. In other cases, it may be that a student-level covariate is the most powerful. That is, fewer schools are needed when a student-level covariate is used. In these cases, a cost analysis should be performed to determine if the additional operational cost associated with acquiring student-level data outweighs the cost of the additional schools required in the study design if relying only on a school-level pretest (Konstantopoulos, 2009).

Conclusions

Our main objectives were to (a) present empirically estimated ICC values for science achievement and compare these values to ICCs for mathematics and reading

achievement, and (b) present empirically estimated R^2 values for covariate sets that are likely to be available to researchers designing CRTs of science interventions. We accomplished these goals using student-level data from the State of Texas.

We estimated unconditional ICC values that range by grade from 0.172 to 0.196 for science achievement in a two-level HLM with students nested in schools. For the three-level HLM, with students nested in schools nested in districts, estimated unconditional ICC values ranged by grade from 0.055 to 0.079 at the district-level, and 0.104 to 0.136 at the school-level. Hence, inclusion of district as a random variable reduces the school-level variance 30-40%.

Our findings also suggest that unconditional ICCs for science achievement are consistently larger than unconditional ICCs for reading. This is an important finding for researchers designing CRTs with science as the primary outcome who try to borrow design parameters from reading as they may run the risk of under-powering their studies. We also found that science ICCs were larger than mathematics ICCs, though the difference was much smaller than that of science and reading.

With respect to the most powerful covariate sets for researchers planning CRTs with science outcomes, grade and lag-length must be considered. In addition, trade-offs in costs associated with including a student-level or school-level covariate will likely enter into discussions. Our results suggest that as expected, when a one-year lagged student-level science pretest is available, it explains the most variance in science achievement at the highest level of nesting. However, this was only available for grade 11. For grades 8 and 10, which have a two-year lagged and three-year lagged science

pretests, respectively, the one-year lagged school-level science pretest was the more powerful covariate set. The one-year lagged school-level science pretest was also the best for grade 5, as no student-level science pretest existed for this grade.

Prior research focused primarily on empirically estimating design parameters for CRTs in mathematics and reading achievement. This study extends the work to science achievement and provides a resource for researchers designing CRTs of science intervention studies. Future research includes extending this work to other states, nationally representative datasets, and other grades in which science is tested.

Limitations

In the TEA dataset, there are two possible outcome variables, the raw score and the scale score, and use of either outcome variable presents limitations. The scale score represents a scaled version of the raw score which takes into consideration the version of the test taken. The scale score is based on all test takers. However, we excluded students taking a non-standard TAKS, special education students, and students receiving a test accommodation from the analysis. The impact of these exclusions cannot be properly modeled using the scale score. Due to the masking of data, use of the raw score outcome limits the generalizability of produced ICCs to only those that are unmasked. This result is undesirable in terms of generalizability, but the raw score is an accurate reflection of student performance for those that took the test. Therefore, in order to compute defensible ICCs that are applicable to students in which data are available we chose to use the raw score as the measure of science achievement.

The masking process makes it difficult to account for specific subpopulations, for example students that received a testing accommodation. In order to assess the impact of our decision to remove students with a testing accommodation, we conducted a sensitivity analysis by also computing ICC values in each year using the two-level and three-level unconditional, demographics-only, and science pretest models, including the unmasked students that received a test accommodation. Although this is not the full population of students that received a test accommodation, the number of unmasked students meeting this requirement varied from a few hundred in 2007 to more than 10,000 in other years. Our results for ICCs were nearly identical across all models tested and years, with the largest difference in ICC between the analysis samples being less than 0.007. This suggests the impact of removing the students that received a test accommodation is small, but that a minor amount of bias is present due to the masking of data in accordance to FERPA. More research into the impact of masking on empirical estimates is warranted, especially as state agencies continue to share data and collaborate with researchers.

References

Bloom, H. S. (2005). Randomizing groups to evaluate place-based programs. In H. S. Bloom (Ed.), *Learning more from social experiments evolving analytic approaches* (pp. 115-172). New York: Russell Sage.

Bloom, H. S., Bos, J. M., & Lee, H. (1999). Using cluster random assignment to measure program impacts: Statistical implications for the evaluation of education programs. *Evaluation Review*, *23*(4), 445-469. doi:10.1177/0193841X9902300405

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

Boruch, R. F., & Foley, E. (2000). The honestly experimental society. In L. Bickman (Ed.), *Validity and social experiments: Donald Campbell's legacy* (pp. 193-239). Thousand Oaks, CA: Sage.

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

Cook, T. D. (2005). Emergent principles for the design, implementation, and analysis of cluster-based experiments in social science. *The Annals of American Academy of Political and Social Science, 599*(1), 176-198. doi:10.1177/0002716205275738

Donner, A., & Klar, N. (2000). *Design and analysis of cluster randomization trials in health research.* London: Arnold.

Donner, A., & Koval, J. J. (1980). The large sample variance of an intraclass correlation. *Biometrika, 67*(3), 719-722. doi:10.1093/biomet/67.3.719

Donner, A., & Koval, J. J. (1982). Design considerations in the estimation of intraclass correlation. *Annals of Human Genetics, 46,* 271-277. doi:10.1111/j.1469-1809.1982.tb00718.x

Fischer, R. A. (1925). *Statistical methods for research workers.* Edinburgh, Scotland: Oliver & Boyd.

Gargani, J., & Cook, T. (2005). *How many schools? Limits of the conventional wisdom about sample size requirements for cluster randomized trials.* Berkeley, CA: University of California, Berkeley.

Hedberg, E. C. (2012). *ICCVAR: Stata module to calculate intraclass correlation (ICC) after xtmixed.* Statistical Software Components. Retrieved from http://ideas.repec.org/c/boc/bocode/s457468.html

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., & Kuyper, A. M. (2012). The variance of intraclass correlations in three- and four-level models. *Educational and Psychological Measurement, 72*(6), 893-909. doi:10.1177/0013164412445193

Institute of Education Sciences. (2013a). *Randomized control trials registry*. Retrieved from http://ies.ed.gov/ncee/wwc/references/registries/RCTSearch/RCTSearch.aspx

Institute of Education Sciences. (2013b). *Search funded research grants and contracts*. Retrieved from http://ies.ed.gov/funding/grantsearch/index.asp

Jacob, R. T., Goddard, R. D., & Kim, E. S. (2014). Assessing the use of aggregate data in the evaluation of school-based interventions: Implications for evaluation research and state policy regarding public-use data. *Educational Evaluation and Policy Analysis, 36*(1), 44-66. doi:10.3102/0162373713485814

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

Kelcey, B., & Phelps, G. (2013). Considerations for designing group randomized trials of professional development with teacher knowledge outcomes. *Educational Evaluation and Policy Analysis, 35*(3), 370-390. doi:10.3102/0162373713482766

Konstantopoulos, S. (2009). Incorporating cost in power analysis for three-level cluster-randomized designs. *Evaluation Review, 33*(4), 335-357. doi:10.1177/0193841X09337991

Mosteller, F., & Boruch, R. (2002). *Evidence matters: Randomized trials in education research.* Washington, DC: Brookings Institution.

Raudenbush, S. W. (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods, 2*(2), 173-185. doi:10.1037/1082-989X.2.2.173

Raudenbush, S. W., & Bryk, A. S. (2002). *Heirarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Raudenbush, S. W., & Liu, X. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological Methods, 5*(2), 199-213. doi:10.1037/1082-989X.5.2.199

Raudenbush, S. W., Martinez, A., & Spybrook, J. (2007). Strategies for improving precision in group-randomized experiments. *Educational Evaluation and Policy Analysis, 29*(1), 5-29. doi:10.3102/0162373707299460

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., Bloom, H., Condon, R., Martinez, A., & Raudenbush, S. W. (2011). *Optimal Design Plus empirical evidence: Documentation for the "Optimal Design" software.* Ann Arbor, MI: University of Michigan.

Spybrook, J. K., & 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. doi:10.3102/01623737093395244

StataCorp. (2011). Stata statistical software: Version 12. College Station, TX: StataCorp LP.

Time4Learning. (2013). *Standardized testing state by state*. Retrieved from http://www.time4learning.com/testprep/

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

CHAPTER IV

INTRACLASS CORRELATIONS FOR THREE-LEVEL MULTI-SITE CLUSTER-RANDOMIZED TRIALS OF SCIENCE ACHIEVEMENT

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).

The applicability of CRTs for studying the effectiveness of educational programs is a result of the inherent nesting that occurs in the educational infrastructure. Students typically learn in traditional classroom environments, and these classrooms are located in schools, which are clustered in districts (Raudenbush & Bryk, 2002). Because educational material is most often delivered through traditional classroom environments, treatment is administered at the cluster level. For example, teachers teach to entire classrooms of students, or new curricula are implemented for whole schools or across districts. However, outcomes are typically measured at the student level through standardized achievement testing. When evaluating the effects of these interventions, a CRT is appropriate because it allows for treatment to be modeled at a different level than the unit of analysis.

Additionally, because in traditional educational environments treatment is administered collectively to groups rather than individually, the standard assumption of

independence is violated. Independence is necessary to operate in an ordinary least squares framework when statistically analyzing experimental data. Using a hierarchical linear structure to model these types of data properly accounts for this violation of independence (Raudenbush & Bryk, 2002).

For reasons of feasibility, a traditional CRT with randomization at the highest level is not always the most practical CRT for an evaluator to utilize. For example, it is unlikely for evaluations of educational interventions to be randomized at the district level; yet, ignoring any district clustering can be detrimental to such studies because the school-level variance will be overestimated, producing a design that is overpowered (Moerbeek, 2004).

One specific type of CRT, a multi-site CRT (MSCRT), is commonly employed in educational research and evaluation studies (Bloom, Richburg-Hayes, & Black, 2007; Spybrook, in press; Spybrook & Raudenbush, 2009). 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. In an educational context, the three-level MSCRT, with district blocks and treatment at the school-level, offers an alternative design to a two-level CRT with students nested in schools, yet still captures the district effect. In this particular three-level MSCRT design, schools are randomly assigned to condition within districts, which act as blocks to remove the district variance that is present.

As a result, MSCRTs offer some benefits over traditional CRTs. For example, MSCRTs are often a more economical design choice, particularly in educational contexts

(Raudenbush & Liu, 2000). Operational costs are typically limited due to the randomization process occurring within sites. Additionally, sample sizes required to achieve a desired power level often are smaller for a three-level MSCRT that randomizes at the school-level than for a traditional two-level CRT that ignores district effects. This occurs, as long as the between-school variance is large, the blocking variable is strong, and there are a large number of schools (Raudenbush, Martinez, & Spybrook, 2007).

A common concern in the design of CRTs that randomly assign treatment at a lower level, like an MSCRT, is the increased likelihood of contamination relative to a traditional CRT (Bloom et al., 2007; Rhoads, 2011; Shadish, Cook, & Campbell, 2002; Schochet, 2008; Torgerson, 2001). For a nested model of students within teachers within schools, a MSCRT would randomly assign classrooms within the same school to either treatment or control. Those not receiving treatment in one classroom may in fact be exposed to treatment via a peer that does receive treatment in a different classroom.

However, contamination need not always be a major concern for MSCRTs in education. For example, in MSCRT designs with students nested within schools nested within districts where the units of randomization (schools within districts) are still geographically separated, the likelihood of one student partially receiving treatment as a result of interaction from another student or between schools would seem low.

Recent empirical research from Doyle and Hickey (2013) on contamination in studies of childhood interventions suggests that the problem does exist, and is discussed frequently in the literature, but actual amounts of contamination are rarely reported in published studies. Thus, the amount of contamination that does exist for MSCRTs in

education is difficult to estimate. However, Rhoads (2011) showed that even when contamination does occur, only very large amounts will decrease the power of a MSCRT design to the amount of a traditional CRT with an equal number of levels.

Planning CRTs and MSCRTs

As in all experimental studies, evaluators must design CRTs with appropriate power to detect an expected effect. Powering a CRT is similar to powering any experimental study in that the evaluator must specify the number of participants required to detect a particular effect-size at an assumed power-level and error-rate. To do this, the evaluator must estimate the amount of variance that exists in the outcome variable, and how much variance is explained by including covariates. For CRTs, which have multiple levels of nesting, the additional stipulation of needing to determine the appropriate sample size at each level of nesting and therefore the amount of variance in the outcome variable that can be partitioned at each level, must be estimated.

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 (school-level), the evaluator must specify the within-site ICC, since the between-site variance is removed by blocking (Konstantopoulos, 2008).

In this study only a natural blocking variable (i.e., district) is considered, thus it is logical to consider a within-district ICC. However, other derived variables (e.g., percentage of students eligible for free or reduced-lunch in a school) can be used as a

blocking variable (Raudenbush et al., 2007). In these cases, one may consider the total variance across schools, prior to blocking, and an estimate of the percent of variance reduced by the blocking.

Empirically estimating ICCs for use in mathematics, reading, and science are a common trend in the education literature (Bloom et al., 2007; Brandon, Harrison, & Lawton, 2013; Hedges & Hedberg, 2007, in press; Jacob, Zhu, & Bloom, 2010; Schochet, 2008; Westine, Spybrook, & Taylor, in press). 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, in press).

Generally, there are two different types of MSCRTs that randomize schools within districts. If the goal of a study is to establish the efficacy of an intervention under ideal conditions, it is likely that there may only be a few (three or less) districts in the study. For example, one district may not provide enough schools to adequately power a study, so similar districts may be recruited in order to increase the sample size. This is similar to an Institute of Education Sciences (IES) goal 3 study, or efficacy trial, in which the evaluator is attempting to ascertain whether a treatment-effect is present, and is not looking to generalize beyond the participants recruited to the study. Often studies of this form will assume fixed district effects (or homogeneous treatment effects), resulting in smaller sample size requirements to achieve necessary power. On the other hand, if the purpose of a study is to establish the effectiveness of an intervention across various

conditions, it is more likely that there will be a large number of districts (eight or more) that are either scattered across a state or across the country. This is similar to an IES goal 4 study, or effectiveness trial, in which the evaluator is interested in the generalizability of an intervention's effect. Typically in this case, the evaluator assumes random district effects (or heterogeneous treatment effect), thereby increasing the sample size requirements in order to achieve necessary power.

Both designs require an estimate of the within-district (school-level) ICC for the power analysis. An estimate of a school-level ICC from a two-level model, which uses all districts across an entire state (or a national sample), is not directly applicable because it is not removing the between-district variation, and in fact will likely overestimate the between-school variance. Furthermore, use of an empirically estimated within-district ICC from a single (typically large) district may not accurately depict the variance structure of a MSCRT design with multiple districts because recruited districts can vary significantly, and may not resemble the source district of the estimate (Hedberg & Hedges, 2011).

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. Conceptually, when districts with more homogenous schools (i.e., having a smaller ICC) are included in the study, fewer schools will be necessary to power a study than when districts with heterogeneous schools (i.e., having a larger ICC) are included. Consider the following example of an MSCRT. Suppose there are two districts which are treated as fixed effects and, hence, the effect size variability is

0. Assuming a significance level of $\alpha = 0.05$, $n = 100$ students per school, and that $r_{L2}^2 = 0.80$ of the variability in achievement using school-level covariates can be explained, then to detect an effect of size of $\delta = 0.20$ it would take $J = 12$ schools per district if the within-district ICC was $\rho = 0.08$, and it would take $J = 18$ schools per district if the within-district ICC was $\rho = 0.16$. Thus, a total of 12 additional schools would be needed for the design with more variance at the school-level.

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. For example, when elements of external validity are prioritized, propensity score based sampling strategies for selecting districts may be used to improve generalization (Tipton et al., 2014). Although these concepts have been put in practice in designs prioritizing internal validity for many years, 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.

Purposes of This Study

In this study, I aim to improve the design of MSCRTs by producing estimates of within-district ICCs for the outcome of science achievement across all districts in an entire state; resulting in a distribution of within-district ICCs. The within-district ICC estimates 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 (Westine et al., in press; Zhu, Jacob, Bloom, & Xu, 2012). Alternatively, evaluators can borrow estimates from another subject, for example mathematics or reading (Hedberg & Hedges, 2011), but borrowing has its limitations, since science estimates appear to be larger than estimates in both these subjects, and recent student-level pretest covariates in other subjects are not always available (Westine et al., in press).

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 by grade in Texas?
2. Does the number of districts in an MSCRT affect the mean within-district ICC?

Method

Data

Student-level 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), which occurs annually in April, student demographic information (e.g., gender, race, socio-economic status), and school and district identifiers. In Texas, as is common in many other states, science is tested only in grades 5, 8, 10, and 11.

The TEA masks data according to the lowest level of clustering. Thus, if fewer than five individuals ($n < 5$) exist in any single demographic group within a cluster (school), achievement scores for every student in that demographic group are masked. Masking occurs more often in elementary grades because school size is generally smaller for elementary grades. For this dataset, the incidence of masked data across grades is, on average, 16%.

The raw score on the TAKS is used as the outcome measure.[9] The analysis is performed using non-masked, non-disabled students with a valid, unique student identification number that took the English version of the TAKS in the standard administration setting that were scored. Records with non-valid or duplicate student identification numbers, as well as students taking an alternate version of the TAKS (i.e., TAKS [Accommodated], TAKS-Modified, TAKS-Alternative) or otherwise requiring an accommodation (i.e., presentation, response, setting, timing and scheduling, braille, large print, and oral administration) are removed from the sample.

Models and ICCs

The primary design examined is the three-level MSCRT with districts treated as sites and schools randomly assigned within sites. Below, I present a model for the three-level MSCRT with treatment at the school level. However, within-district ICCs are estimated using an unconditional two-level HLM for each individual district. The net result is a distribution of ICC values across districts. After the three-level MSCRT, I present the model for a two-level CRT, which is used to empirically estimate an ICC value for each district.

The three-level MSCRT. The unconditional three-level MSCRT with students (Level 1) nested within schools (Level 2) nested within districts (Level 3), where

[9] Alternatively, the scale score could have been used, which represents a scaled version of the raw score. The scale score takes into consideration the version of the test taken, and is derived from outcomes across all versions of the TAKS. Given that data from non-standard TAKS are excluded from the analysis, the use of the raw score limits any bias in the results to that of masking as opposed to also introducing bias through scaling that is attributable to the test version.

blocking occurs on districts and treatment is administered at the school-level, is as follows. The Level 1 or student-level model is:

$$Y_{ijk} = \pi_{0jk} + e_{ijk} \qquad e_{ijk} \sim N(0, \sigma^2), \qquad \textbf{[36]}$$

where Y_{ijk} is the outcome for individual $i \in \{1, \ldots, n_{jk}\}$, in school $j \in \{1, \ldots, J_k\}$, in district $k \in \{1, \ldots, K\}$, π_{0jk} is the mean achievement at school *j* in district *k*, and e_{ijk} is a random student effect, which is assumed to be normally distributed with a mean of 0 and homogenous variance σ^2. Therefore, σ^2 is the variance in achievement among students within schools. The Level 2 or school-level model is:

$$\pi_{0jk} = \beta_{00k} + \beta_{01k} W_{jk} + r_{0jk} \qquad r_{0jk} \sim N(0, \tau_\pi), \qquad \textbf{[37]}$$

where β_{00k} is the mean in district *k*, W_{jk} is the school-level treatment indicator typically set to 0.5 for the treatment group and -0.5 for the control group, β_{01k} is the mean effect of treatment for the k^{th} district, and r_{0jk} is the random school effect, which is assumed to be normally distributed with a mean of 0 and homogenous variance τ_π. Therefore, τ_π is the variance in mean achievement among schools within districts. The Level 3 or district-level model with heterogeneous treatment effects (or random district-effects) is:

$$\beta_{00k} = \gamma_{000} + u_{00k} \qquad u_{00k} \sim N\left(0, \tau_{\beta_{00}}\right), \qquad \textbf{[38]}$$

$$\beta_{01k} = \gamma_{010} + u_{01k} \qquad u_{01k} \sim N\left(0, \tau_{\beta_{11}}\right),$$

$$cov(u_{00k}, u_{01k}) = \tau_{\beta_{01}}$$

where γ_{000} is the grand mean, u_{00k} is a random district effect which is assumed to be normally distributed with a mean of 0 and variance $\tau_{\beta_{00}}$, γ_{010} is mean effect of treatment, and u_{01k} is a random treatment effect which is assumed to be normally distributed with a

mean of 0 and variance $\tau_{\beta_{11}}$. Therefore $\tau_{\beta_{00}}$ is the variance in mean achievement among districts, $\tau_{\beta_{11}}$ is the variance in the treatment effect among districts, and $\tau_{\beta_{01}}$ is the covariance between district-specific mean achievement and treatment effects.

If homogeneous treatment effects (or fixed district-effects) are assumed, then $\tau_{\beta_{11}} = 0$, and u_{01k} becomes a fixed-effect for each district. Additionally, $cov(u_{00k}, u_{01k}) = 0$.

For models with random district-effects or fixed district-effects, the parameters are standardized so that there is only one[10] ICC value,

$$\rho = \frac{\tau_\pi}{\tau_\pi + \sigma^2}. \qquad \textbf{[39]}$$

The unconditional two-level model. 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:

$$Y_{ij} = \beta_{0j} + r_{ij} \qquad r_{ij} \sim N(0, \sigma^2), \qquad \textbf{[40]}$$

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:

---

[10] Other parameterizations include two ICC values for a three-level MSCRT with treatment at Level 2, see, for example, Spybrook, Hedges, and Borenstein (in press).

$$\beta_{0j} = \gamma_{00} + u_{0j} \qquad u_{0j} \sim N(0, \tau_{00}), \qquad \textbf{[41]}$$

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,

$$\rho = \frac{\tau_{00}}{\tau_{00} + \sigma^2}. \qquad \textbf{[42]}$$

It is important to present both the three-level MSCRT and two-level CRT models because ICCs from the two-level CRT are used to inform the design of the three-level MSCRT. Although equations for the ICCs in each model look similar, they are actually quite different. In a MSCRT, the variance among districts is accounted for by blocking on districts and [39] represents the within-district variance among schools. In equation [42] variance among districts is not accounted for. However, it is also not present in our context because the model is only used one district at a time. To power an MSCRT, an estimate of the within-district variance among participating districts is needed. This is accomplished, for example, by taking the mean school-level ICC for these specific districts. In this way, the ICC calculations using a two-level model for each district inform the ICC estimate for a MSCRT.

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, 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. Here the interest is whether the number of districts in the design affects the within-district ICC estimate. Thus, I generate and compare confidence intervals on the mean within-district ICC for MSCRTs of different sizes. The analysis is limited to Grades 5, 8, 10, and 11 because these are the grades in which science is tested in Texas. However, in higher grades the number of schools per district is smaller, and so there is less flexibility in designing studies due to sample size limitations. The analysis is also limited to balanced MSCRTs where the average number of students per school is chosen to include the vast majority of schools, where $n \geq 25$.

First, the set of districts in Texas that can feasibly be used when conducting MSCRTs in each grade is defined. Two broader classes of MSCRTs that are commonly used in the education literature, those with only a few districts, and those with many districts, are of interest. For a MSCRT with only a few districts, a large number of schools per district are needed to adequately power the study. For a MSCRT with many districts, a small number of schools per district are needed. To operationalize a MSCRT with only a few districts I use $K = 3$ districts with a corresponding value of $J \geq 20$ schools per district. To operationalize a MSCRT with many districts, $K = 10$ with a

corresponding value of $6 \leq J < 20$ is used. The number of schools per district is used to identify the individual districts as either (1) eligible for a MSCRT with only a few districts; (2) eligible for a MSCRT with many districts; or (3) not eligible for a MSCRT (i.e., too small). The three categories of districts are mutually exclusive.

The choice for the number of districts and schools per district for each case was based on two primary decisions. First, I wanted to be consistent with the empirical literature, and these choices correspond to existing examples of three-level and four-level MSCRTs (excluding matched-pairs designs) taken from the What Works Clearing House for Goal 3 and Goal 4 studies (Spybrook, in press), where Level-2 corresponds to school. In these examples, for studies with only a few districts the corresponding *J* value ranges between 19 and 24. For studies with a larger number of districts the corresponding *J* value is between four and six. While sample sizes from actual studies vary and depend upon a number of circumstances (e.g., desired minimum detectible effect-size, estimated ICC value, cost structure), the sample sizes chosen are seemingly representative of the two broader classes of MSCRTs. Second, a consistent number of total schools in the two designs is maintained in order to investigate differences that may exist between the two designs. In each MSCRT, the total number of schools is 60. Ultimately, the district, school, and student sample size requirements for each of the cases limits the number of districts eligible for each design.

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. I conduct a *t*-test for each grade to compare the mean within-district ICC for designs with only a few

districts to designs with many districts. More specifically, grade is tested at the $\alpha = 0.0125$ significance level (thereby accounting for multiple tests), to determine 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 $H_0: \bar{\rho}_{25,20,3} - \bar{\rho}_{25,6,10} = 0$ and $H_A: \bar{\rho}_{25,20,3} - \bar{\rho}_{25,6,10} \neq 0$.

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 4.1 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. Given that ICCs are limited on the range [0, 1], with most examples in the educational literature emphasizing 0.1 to 0.3, these distributions are expected to be skewed. In fact, the findings for science are similar to those found for mathematics and reading (Hedberg & Hedges, 2011).

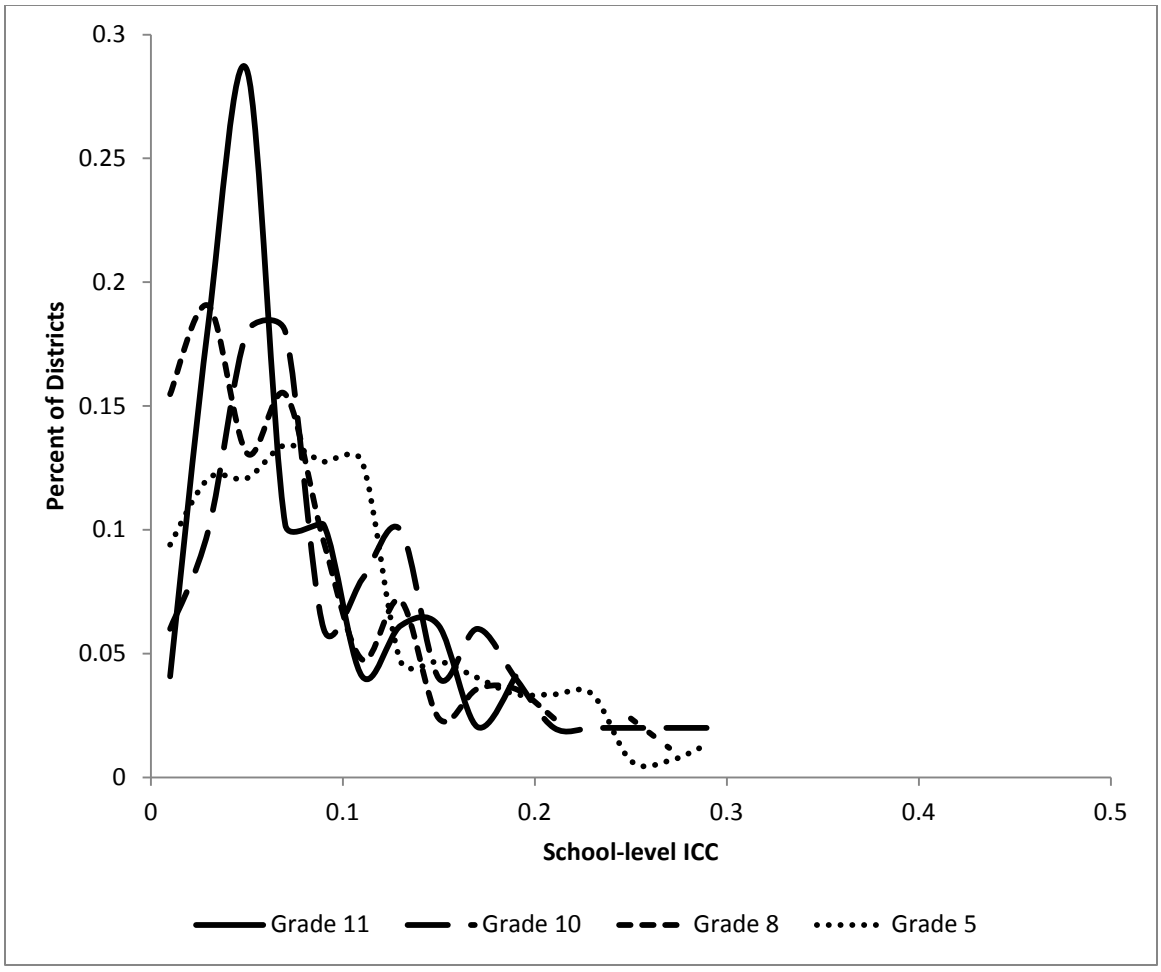


Figure 4.1. Distribution of unconditional school-level ICCs in science by grade for districts with four or more schools in Texas.

The distributions are fairly consistent across grades as well, as can be seen in Table 4.1. 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.

Table 4.1

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		

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 4.2 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, $\bar{\rho}_{25,6,10} = 0.0843$ ($SE = 0.0069$).

Table 4.2

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*	*df.*	*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*

In Grade 8, no significant difference in the mean within-district ICC for the two designs is found ($p = 0.5196$); however, the ability to determine a difference is impacted by having only four eligible districts for the design with fewer districts. For the design with only a few districts, I find $\bar{\rho}_{25,20,3} = 0.1102$ ($SE = 0.0300$), but for the design with many districts, $\bar{\rho}_{25,6,10} = 0.0877$ ($SE = 0.0096$).

In Grades 10 and 11, there are not enough large districts to execute a balanced MSCRT with $K = 3$ districts, $J \geq 20$ schools per district, and $n \geq 25$ students per school because there are only two districts per grade that meet these criteria. Hence, mean within-district ICCs for the two different MSCRT designs cannot be compared. Thus, for each grade only the mean within-district ICC for the design with many districts is

presented. In Grade 10, I find $\bar{\rho}_{25,6,10} = 0.0957$ ($SE = 0.0153$), and in Grade 11, I find $\bar{\rho}_{25,6,10} = 0.0893$ ($SE = 0.0135$).

Discussion

Efforts to increase the rigor of science, technology, engineering, and mathematics (STEM) educational evaluations have emphasized designs such as CRTs which account for variance found in different levels of nesting. Part of the job of STEM evaluators is to efficiently design studies including appropriately powering designs by accurately estimating variance at each level of nesting. In cases where treatment diffusion is of little concern, such as whole school interventions, MSCRTs with schools randomly assigned within districts are often an efficient design choice. While empirical estimates of design parameters for CRTs in STEM continue to appear in the literature, they have primarily focused on outcomes of mathematics. Empirical estimates of ICCs for studies of science achievement outcomes are rare, and have yet to consider MSCRT designs. This study fills this void in the empirical literature.

Using student achievement data from Texas, the distribution of within-district ICCs required for powering an MSCRT of science achievement primarily exist between 0 and 0.30. Also, the mean within-district ICC does not vary much by grade (0.0781-0.0982). These estimates are much smaller than school-level ICC estimates from a two-level model, using statewide data, which ranged from 0.172-0.196 (Westine et al., in press). While average within-district ICC estimates by grade are expected to be smaller than those from a two-level model because district variance is not accounted for, the

actual difference (approximately 50%) highlights the importance of recognizing the conceptual difference between the two school-level ICC values when selecting parameter values for a power analysis. In the present study, the two-level model is used only for individual districts, thus eliminating any district variance that would exist when multiple districts are present, but not accounted for by the model.

The two common MSCRT scenarios by grade demonstrate that ICC estimates for MSCRTs can be refined further in some cases. By basing this investigation in an empirical example, I show that two common design types drastically, but uniquely limit the eligibility of districts for each design by grade. For example, in Grade 5, only about one-third of districts have school-level sample sizes large enough ($J \geq 20$) to participate in a MSCRT design that utilizes only a few districts ($K = 3$). Another one-third are in mid-size districts ($6 \leq J < 20$), which are used in MSCRT designs with many districts. The final one-third cannot reasonably be used in a balanced MSCRT because they are too small.

In Texas, only Grade 5 had enough districts and schools to meaningfully compare means for district subsets defined by the type of MSCRT; in higher grades, not enough (or barely enough) districts were eligible for a MSCRT with $K = 3$ districts. For Grade 5, 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. In other states where the structure of schools in districts is different, it will be useful to explore whether significant differences exist between mean within-district ICC values in higher grades.

Using Optimal Design, I demonstrate the impact of the different ICC values for Grade 5. Suppose one desires to power a MSCRT with $K = 3$ districts in order to detect an effect size of $\delta = 0.20$. Assuming a significance level of $\alpha = 0.05$, $n = 25$ students per school, $J = 20$ schools per district, a school-level covariate explains 60% of the variability in achievement, and the within-district ICC is $\rho = 0.084$, the mean value for mid-sized districts, then the study would be sufficiently powered at 0.81. However, if $\rho = 0.130$, the mean value for large districts, the study would be underpowered at 0.72.

Thus, 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. Larger districts, which are commonly used in IES goal 3 or similar-type studies, seem to be associated with larger within-district ICC values, which unfortunately is counterproductive to reducing costs. Specific district ICC values can vary considerably, though, so caution should be exercised in estimating this value.

In IES goal 3 or similar-type studies where generalizability is not emphasized, evaluators would benefit from documenting within-district ICC values, or predictive-type models which could help identify districts with smaller school-level ICC values. For example, for Grade 5 in Texas, the three large districts with the smallest within-district ICCs have an average ICC of only 0.0442. Considering the example above, the power of the study using these particular districts is 0.8946. This suggests fewer participating schools per district are actually needed to meet an acceptable level of power.

MSCRT designs including many districts, such as IES goal 4 or similar-type designs, offer more flexibility in the choice of districts. In Grade 5, where there are many

large districts to choose from, a design would seemingly also benefit from the evaluator targeting recruitment of districts with the smallest school-level ICCs. However, selecting specific districts with small ICC values could be challenging as two important criteria for designs, generalizability and efficiency, do not necessarily work in harmony.

As demonstrated in this paper, certain districts align better with one type of MSCRT than another. To further improve the design of MSCRTs of science achievement, emphasis should be placed on expanding empirical estimates of within-district ICCs across states in order to test the properties of within-district ICCs from different educational structures.

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

Doyle, O., & Hickey, C. (2013). The challenges of contamination in evaluations of childhood interventions. *Evaluation, 19*(2), 183-194. doi:10.1177/1356389013482610

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. (in press). Intraclass correlations and covariate outcome correlations for planning 2 and 3 level cluster randomized experiments in education. *Evaluation Review*.

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

Moerbeek, M. (2004). The consequence of ignoring a level of nesting in multilevel analysis. *Multivariate Behavioral Research, 39*(1), 129-149. doi:10.1207/s15327906mbr3901_5

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

Raudenbush, S. W., & Liu, X. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological Methods, 5*(2), 199-213. doi:10.1037/1082-989X.5.2.199

Raudenbush, S. W., Martinez, A., & Spybrook, J. (2007). Strategies for improving precision in group-randomized experiments. *Educational Evaluation and Policy Analysis, 29*(1), 5-29. doi:10.3102/0162373707299460

Rhoads, C. H. (2011). The implications of "contamination" for experimental design in education. *Journal of Educational and Behavioral Statistics, 36*(1), 76-104. doi:10.3102/1076998610379133

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

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference.* Boston, MA: Houghton Mifflin Company.

Spybrook, J. K. (in press). Detecting intervention effects across context: An examination of the precision of cluster randomized trials. *Journal of Experimental Education*. doi:10.1080/00220973.2013.813364

Spybrook, J. K., & 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

Spybrook, J. K., Hedges, L. V., & Borenstein, M. (in press). Understanding statistical power in cluster-randomized trials: Challenges posed by differences in notation and terminology. *Journal of Research on Educatcational Effectiveness*.

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

Tipton, E., Hedges, L., Vaden-Kiernan, M., Borman, G., Sullivan, K., & Caverly, S. (2014). Sample selection in randomized experiments: A new method using propensity score stratified scoring. *Journal of Research on Educational Effectiveness, 7*(1), 114-135. doi:10.1080/19345747.2013.831154

Torgerson, D. J. (2001, February 10). Contamination in trials: is cluster randomisation the answer? *British Medical Journal, 322*, 355-357. doi:10.1136/bmj.322.7282.355

Westine, C. D., Spybrook, J. K., & Taylor, J. A. (in press). An empirical investigation of variance design parameters for planning cluster-randomized trials of science achievement. *Evaluation Review*.

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

CHAPTER V

CONCLUSION

In this final chapter, I first summarize and review the main findings from each of the essays in order to draw conclusions regarding the impact of the research on improving the design of cluster-randomized trials in education. Next, I collectively discuss the limitations of the essays. Finally, I consider where the research can go from here, and highlight what I believe to be productive avenues for future research.

Summary and Review of Main Findings

Throughout the essays, there are several specific noteworthy contributions. First, in Chapter II, the analysis of relative efficiency in a three-level CRT shows that both Level 2 and Level 3 ICCs are reasonably robust to misspecification, even when both ICCs are incorrect. This finding mirrors foundational work in this area by Korendijk, Moerbeek, and Maas (2010). This is useful for science education evaluators who face significant uncertainty when specifying parameters, because small increases in sample size can compensate for misspecification.

Chapters III and IV contribute to the formation of an empirical base of design parameters for the evaluation of science education interventions using CRT and MSCRT designs. Empirical estimates of design parameters for science education do not currently exist across the range of grades and formats in which science is tested, and the results of

this dissertation fill this void for traditional three-level CRTs as well as MSCRTs. The results from the one state considered, Texas, show that in all grades in which science is tested, science ICCs are equal to or larger than both mathematics and reading, and that there is a much larger difference between science and reading ICCs than between science and mathematics ICCs. Additionally, within-district ICCs for science were typically found to range between 0 and 0.30. However, in grade 5, the average within-district ICC varies according to the number of districts used in a design. By considering actual data from Texas in two commonly used MSCRT designs, within-district ICCs were shown to be larger for designs with only a few districts and a large number of schools than for designs that have many districts and only a few schools in grade 5. On average, within-district ICCs are approximately half the size of a school-level ICC from a two-level model with students nested in schools.

Finally, Chapter III is the first study to explore the hierarchy of educational pretest covariates for a specific application. Beyond specific findings for science, the analysis likely will serve as an example for other researchers as empirical estimates of design parameters are explored in new subject areas. While others studies have presented R^2 estimates for various covariate models (Bloom, Richburg-Hayes, & Black, 2007; Hedges & Hedberg, 2007), the context has always been mathematics or reading achievement where pretest covariates are readily available. In the context of science, where annual testing is not the norm, the comparison of lagged pretests is relevant and important. Results of this study show that when available a one-year lagged student-level science pretest is the best predictor of science achievement, although a one-year school-

level science pretest is a very good predictor too. When the one-year lagged student-level science pretest is unavailable, which is true in all grades except grade 11, the one-year lagged school-level science pretest is preferred.

Other contributions are less specific, but are significant because they are timely. For example, as educational evaluations increasingly involve more than two levels of nesting, the need for design parameters values estimates from three-level models is especially relevant. Additionally, the notion of improved outcomes in science, technology, engineering, and mathematics (STEM) disciplines continues to be relevant to educational policy makers. The tackling of research questions pertinent to science education evaluation methods helps to ensure science education evaluations are of the highest quality. Finally, evaluation efficiency is a topic that is touched on in each of the three essays. The dissertation serves as an important example for evaluation practitioners of how evaluative decisions like the specification of ICC values can impact the cost-effectiveness of an evaluation.

Overall, the results of this dissertation are encouraging. While it is important for evaluators to be cost-effective in their evaluation designs, three-level CRT designs are reasonably robust to misspecification of one, two, or both ICCs.

Prior to this research effort, science education evaluators planning a three-level CRT had very little guidance regarding how to appropriate estimate ICC and R^2 values. Decisions regarding sample sizes need to be justified, and there was very literature to reference for defensible claims regarding variance. At best, evaluators needed to rely on

borrowing ICC values from mathematics or reading, and attempt to justify their applicability.

The results of this research should instill confidence in science education evaluators. Science education evaluators are now equipped with useful information to accurately design and better plan more efficient studies.

Limitations

In this section, I describe some limitations of the research, including concerns regarding the masking of data, assumptions, and the overall generalizability of the findings. First, because of the large population in Texas, sample sizes for ICC and R^2 estimates are large, and help to create precise estimates. However, the data used in Chapter III and IV were subject to masking. Thus, despite the large dataset, some student records were not available for analysis, which introduces an unknown amount of bias in the results. However, many of the student achievement scores were masked because the student took a different version of the TAKS test, and therefore would not have been considered in the analysis anyway. This diminishes the impact of the masking. Nevertheless, in any research effort it is important to adequately describe the population served by the intervention, and in this research I was unable to consider the full population of students in Texas.

Throughout the dissertation, assumptions were made to clarify or sometimes simplify the analysis. For example, models in the analysis are assumed to be balanced, when in all likelihood a balanced design is unrealistic. In other instances, assumptions were made to consider a single representative case because it would be impractical to

explore all alternatives. In these situations, I do not presume the representative case represents all possible configurations of parameters. Results may different depending on the assumptions made. Therefore, the precision of estimates is biased due to these assumptions. The extent of this bias is likely small, unless significant imbalance occurs; the actual amount of bias is unknown.

While the estimates of ICCs for CRTs and MSCRTs should be considered when designing a CRT or MSCRT in another state, it is unlikely that the educational structure in other states will align with those in Texas. Data for other states and grades are needed in order to assess the impact of various cleaning and analysis steps. Hence, the applicability of the results to be generalized beyond population of students in Texas is limited.

New Directions

To conclude, I pause to reflect on the motivation for this dissertation, and consider extensions of this work that are most relevant going forward. When I originally conceptualized this dissertation, I was working under motivation to equip science educational evaluators with the tools necessary to accurately design a three-level cluster-randomized trial with students nested in schools nested in districts. Recent work in mathematics and reading had expanded the repertoire of design parameters to include ICCs and R^2 values that could be used to design three- or even four-level models, but science education evaluators still were without access to empirical estimates specific to their discipline.

Of course, there are multiple directions for future research in this area. One obvious need is to extend the analysis to four levels by including the teacher level. While Zhu, Jacob, Bloom, and Xu (2012) find that, in most cases, a three-level analysis can be performed using two levels, ignoring the classroom level, and therefore the classroom level is less important, many interventions target the teacher level for assignment. For example, technological interventions like digital instructional delivery may more easily be implemented at the classroom level, and the threat of contamination is low because students do not interact across classrooms during the actual instructional delivery.

Adding a fourth level of data is also helpful for considering outcomes associated with teacher professional development. Kelcey and Phelps (2013) were the first to provide empirical estimates of design parameters for teacher professional development, but the outcomes did not include science education. Because there is such a demand for improved STEM teaching and learning, adding a teacher level will help researchers to better explore the links between science teacher professional development and student achievement.

One challenge with adding a fourth level is that the masking process will most assuredly render the data unusable. As state agencies increasingly partner with researchers, steps need to be taken to avoid the masking process, yet still maintain an adherence to privacy-related issues. By working together to understand the nuances of the data, researchers and data managers can improve the quality of CRTs and other evaluation methodologies in education.

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

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

Kelcey, B., & Phelps, G. (2013). Considerations for designing group randomized trials of professional development with teacher knowledge outcomes. *Educational Evaluation and Policy Analysis, 35*(3), 370-390. doi:10.3102/0162373713482766

Korendijk, E. J., Moerbeek, M., & Maas, C. J. (2010). The robustness of designs for trials with nested data against incorrect initial intracluster correlation coefficient estimates. *Journal of Educational and Behavioral Statistics*, 566-585. doi:10.3102/1076998609360774

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 A

Average R^2 Values for Models with Demographics and Pretest Covariates

Table A

Average R^2 for a Two-Level and Three-Level HLM of Science Achievement with Student-Level Demographics and the Most Recent Student-Level Science, Reading, or Mathematics Pretest Covariate by Grade, 2008-2011

Pretest Subject	Grade	Two-Level HLM		Three-Level HLM		
		R^2_{L1}	R^2_{L2}	R^2_{L1}	R^2_{L2}	R^2_{L3}
Science	5	-	-	-	-	-
	8	0.335	0.705	0.335	0.687	0.679
	10	0.485	0.820	0.485	0.825	0.789
	11	0.519	0.913	0.519	0.908	0.925
Reading	5	0.311	0.677	0.311	0.623	0.740
	8	0.371	0.776	0.371	0.785	0.678
	10	0.245	0.780	0.245	0.803	0.725
	11	0.267	0.786	0.267	0.789	0.789
Mathematics	5	0.313	0.701	0.313	0.626	0.783
	8	0.453	0.800	0.453	0.770	0.790
	10	0.471	0.874	0.471	0.886	0.816
	11	0.479	0.861	0.479	0.865	0.843

Table B

Average R^2 for a Two-Level and Three-Level HLM of Science Achievement with Student-Level Demographics and a One-Year School-Level Science, Reading, or Mathematics Pretest Covariate by Grade, 2008-2011

Pretest Subject	Grade	Two-Level HLM		Three-Level HLM		
		R^2_{L1}	R^2_{L2}	R^2_{L1}	R^2_{L2}	R^2_{L3}
Science	5	0.101	0.731	0.101	0.651	0.898
	8	0.129	0.819	0.129	0.775	0.860
	10	0.129	0.877	0.129	0.884	0.822
	11	0.131	0.874	0.131	0.880	0.844
Reading	5	0.101	0.637	0.101	0.587	0.705
	8	0.129	0.717	0.129	0.713	0.658
	10	0.129	0.757	0.129	0.778	0.680
	11	0.131	0.736	0.131	0.770	0.706
Mathematics	5	0.101	0.675	0.101	0.607	0.772
	8	0.129	0.757	0.129	0.737	0.720
	10	0.129	0.834	0.129	0.847	0.763
	11	0.131	0.838	0.131	0.848	0.796

Appendix B

Human Subjects Institutional Review Board Approval Letter

WESTERN MICHIGAN UNIVERSITY
Human Subjects Institutional Review Board

Date: April 15, 2013

To: Jessaca Spybrook, Principal Investigator
Carl Westine, Student Investigator for dissertation

From: Amy Naugle, Ph.D., Chair

Re: HSIRB Project Number 13-04-10

This letter will serve as confirmation that your research project titled "Improving the Design of Cluster Randomized Trials: Three Essays to Inform the Selection of Design Parameter Values" has been **approved** under the **exempt** category of review by the Human Subjects Institutional Review Board. The conditions and duration of this approval are specified in the Policies of Western Michigan University. You may now begin to implement the research as described in the application.

Please note: This research may **only** be conducted exactly in the form it was approved. You must seek specific board approval for any changes in this project (e.g., ***you must request a post approval change to enroll subjects beyond the number stated in your application under "Number of subjects you want to complete the study***)." Failure to obtain approval for changes will result in a protocol deviation. In addition, if there are any unanticipated adverse reactions or unanticipated events associated with the conduct of this research, you should immediately suspend the project and contact the Chair of the HSIRB for consultation.

Reapproval of the project is required if it extends beyond the termination date stated below.

The Board wishes you success in the pursuit of your research goals.

Approval Termination: April 15, 2014

Walwood Hall, Kalamazoo, MI 49008-5456
PHONE: (269) 387-8293 FAX: (269) 387-8276