4-2019

Design Parameters for Planning Cluster Randomized Trials of Cognitive Skill Interventions: An Empirical Analysis Using the Collegiate Learning Assessment

Yu Du
Western Michigan University, cunmeidu@hotmail.com

DESIGN PARAMETERS FOR PLANNING CLUSTER RANDOMIZED TRIALS OF COGNITIVE SKILL INTERVENTIONS: AN EMPIRICAL ANALYSIS USING THE COLLEGIATE LEARNING ASSESSMENT

by

Yu Du

A dissertation submitted to the Graduate College
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
Educational Leadership, Research, and Technology
Western Michigan University
April 2019

Doctoral Committee:
Jessaca Spybrook, Ph.D., Chair
Gary Miron, Ph.D.
David Reinhold, Ph.D.

ACKNOWLEDGEMENTS

This journey would not have been possible without the support of the Council for Aid to Education (CAE), my committee, professors, staff, and friends. First, I would like to express my gratitude to my data provider for my dissertation, Collegiate Learning Assessment (CLA) test administrator CAE, which enabled me to conduct the research.

Next, I would like to express my profound gratitude to my dissertation committee, Dr. Jessaca Spybrook (Chair), Dr. Gary Miron, and Dr. David Reinhold. I am indebted for their time, efforts, and contributions serving on this committee. And I thank them for their valuable and constructive feedback which definitely enhanced the quality of my dissertation. Special mention goes to my enthusiastic advisor, Dr. Spybrook. Her role as my advisor and the committee chair was critical in this process. I thank her wholeheartedly, not only for helping me explore ideas but also for her scholarly input, tremendous academic support, and for giving me the opportunity of trying things in unique ways.

Similarly, my gratitude goes to Dr. Ming Li, Dean of the College of Education and Human Development; Dr. Donna Talbot, Department Chair of Educational Leadership, Research, and Technology (ELRT); and Dr. Brooks Applegate, Director and Professor of the Evaluation and Measurement (EMR) program. I am particularly indebted to Dr. Nancy Mansberger, Director of Academic Labor Relations, and Dr. Jianping Shen, who have been truly dedicated mentors and generous supporters during my doctoral studies here. And I gained a lot from them,

through their personal and scholarly interactions and their suggestions at various points of my research experience.

I would like to express my very appreciation to faculty and staff of ELRT for the substantial influence that their courses, practicum, and research projects have had on my research and I do hereby acknowledge all of them. They are: Dr. Patricia Reeves, Professor of Evaluation, Measurement, and Research; Dr. Chris Coryn, Director of Interdisciplinary Ph.D. in Evaluation; Dr. Andrea Beach and Dr. Wanda Hadley. My appreciation also goes to Mary Ebejer, the writing coach (ELRT) and Kim Ballard (Writing Center) for their advising on dissertation writing. Moreover, my appreciation for Lorie Wolfe and Diadrian Washington (Tate Grant and Innovation Center) for their concerns and wishes. For our staff members, Diane Bourgeois, Linda Root, Gretchen St. John, and Linda Walters, I appreciate their wonderful services all these years.

Then, I am also hugely appreciative for the academic engagement and research experience in fruitful collaboration with all the following personnel. Special thanks to Dr. Susan Stapleton, Dean of the Graduate College; Dr. Christine Byrd-Jacobs, Associate Dean of the Graduate College; Dr. Marianne Di Pierro (retired); Carson Leftwich (retired); Dr. Julie Rickert (formatting editor); Julia Mays, Associate Director Research Compliance; Dr. Jody Brylinsky (retired), Vice Provost of Institutional Effectiveness; Kathleen Springsteen, Assistant Director of Integrated Program Review and Accreditation; Dr. Fen Yu, Director of Institutional Research; Jason Jach, Associate Director of Institutional Research; William McQuitty, Institutional Research Analyst; Karen Stokes Chapo, Program Manager Assessment; Dace Copeland, Business Manager Senior in Academic Affairs; Dr. Ying Zeng, Director of Asian Program; Dr.

Xiaojun Wang, Department of World Languages; Dr. Dan Fitzpatrick, Dr. Daniel Collier, and Dr. Mary Anne Sydlik from CRICPE (Center for Research on Instructional Change in Post-Secondary Education). As time goes on, I realize clearly the huge impact that has on my direction for my academic career.

To my beloved parents and my older brother, I am deeply thankful for their love, emotional and financial support, and sacrifices. I owe a great debt of gratitude to them. I always knew that they believed in me and wanted the best for me and gave me the freedom to pursue my dream. Additionally, many thanks to my aunt, uncles, and cousins both in China and USA for all their efforts in helping me to achieve my goal.

Last but not least, I would like to thank the WMU Graduate Student Association (GSA) for the fellowship. I would like to thank my lifetime friends in China, Weidong Jiang, Pingzhang ,and Dan Zhou for having being my side. I also harvested friendship here with Emily, Shannon, Rita, Shanon, Katrina, Jane and Clinton, my host family. In addition, thank my past and current classmates' academic support, they are: Jiangan, Zhenji, Huang Wu, Mitsuyo, Wesssam, Laura, Fanny, Jen, Bo, Ran Shi, Qian Wang, Mustafa, Diayana, Francis, and Dustin. Thank all my friends and classmates for the great times that we have shared and rapport that we forged over these years.

Yu Du

DESIGN PARAMETERS FOR PLANNING CLUSTER RANDOMIZED TRIALS OF COGNITIVE SKILL INTERVENTIONS: AN EMPIRICAL ANALYSIS USING THE COLLEGIATE LEARNING ASSESSMENT

Yu Du, Ph.D.

Western Michigan University, 2019

Recently, higher education has started to place a premium on rigorous research that uses randomized controlled trials (RCTs) to test the impact of educational interventions. This may be due in part to concerns about a deficiency of high-quality evidence of the effectiveness of programs, policies, and practices to improve undergraduate students’ outcomes. Given the naturally nested structure in higher education, e.g., students nested in colleges/universities, researchers in higher education start considering a specific type of RCT called a cluster randomized trial (CRT), which have been frequently used in K-12 impact research. In a CRT, whole clusters, such as colleges/universities, are assigned to treatment or control conditions. Just like in RCTs, it is critical that CRTs are designed with adequate power to detect a meaningful treatment effect. However, the multilevel nature of CRTs makes the power analyses more complex than in a RCT. Two key design parameters that are necessary in order to calculate the power for a CRT are the intraclass correlation coefficient (ICCs), or the percent of variance in the outcome that is between clusters, and the variance in the outcome that is explained by covariates (R^2). So far, a rich body of evidence of empirical estimates of these design parameters is available in K-12 settings. However, these design parameters are context-specific and there is a lack of empirical evidence of estimates of these design parameters in high education settings.

The purpose of this study is to empirically estimate ICCs and R^2 values for planning CRTs aimed at evaluating the efficacy of collegiate cognitive skills interventions in higher education. This study uses data from the Collegiate Learning Assessment (CLA), which is a standardized test measuring students' cognitive ability in higher education. A series of two-level hierarchical linear models were employed to calculate the design parameters. The unconditional model, or model with no covariates, was used to calculate the ICCs. Models with student level and school level covariates were then used in order to calculate the R^2 values. The influence of these design parameters on statistical power was examined by calculating the minimum detectable effect size under various sample sizes using the estimated design parameters.

Across all samples and outcomes, the ICC estimates ranged from 0.194 to 0.353. That is, between 19 and 35 percent of the variance in test scores was between colleges/universities. The proxy variables for the student level pretest and school level pretest had the greatest explanatory power of the covariates considered and in most cases explained between 60 and 86 percent of the between school variance in the outcomes. This suggests that including a proxy for pretest, either at the student or school level, is critical in designing a CRT as it will greatly increase the statistical power of the study to detect a meaningful effect. The empirical estimates of design parameters in this study represent the beginning of a collection of design parameters relevant for those planning CRTs to test interventions in higher education and extending this work to other outcome domains in higher education would be useful.

TABLE OF CONTENTS

CHAPTER

CHAPTER

Table of Contents—Continued

CHAPTER

CHAPTER

Table of Contents—Continued

LIST OF TABLES

List of Tables—Continued

LIST OF FIGURES

LIST OF ABBREVIATIONS

ANOVA	Analysis of variance
ASAP	Accelerated Study in Associates Program
AW	Analytic Writing Task
CRUE	Center for Research on Undergraduate Education
CCC	Connected Chemistry Curriculum
CCTT	Cornell Critical Thinking Tests
CCTST	California Critical Thinking Skills Test
CSR	Class Size Reduction
CSRD	Comprehensive School Reform Demonstration
CUNY	City University of New York
ERIC	Education Resources Information Center
ESI	Experimental Sites Initiative
FAFSA	Free Application for Federal Student Aid
FTE	Full Time Equivalent
FITW	First in the World
FYE	First-year Experience
HBCU	Historically Black College or University
HD	Hierarchical Design
HEA	Higher Education Act
HEOA	Higher Education Opportunities Act
MDRC	Manpower Demonstration Research Corporation
MCAR	Missing Completely at Random
MAR	Missing at Random
MI	Multiple Imputation
MNAR	Missing Not at Random
NCER	National Center for Education Research
NCLB	No Child Left Behind Act
NCSER	National Center for Special Education Research
NRC	National Research Council
PSAE	Postsecondary and Adult Education
PT	Performance Task
SIP	Strengthening Institutions Program
STAR	Student/Teacher Achievement Ratio
VIF	Variance Inflation Factor
WGCTA	Watson–Glaser Critical Thinking Appraisal

CHAPTER I

INTRODUCTION

A randomized controlled trial (RCT), also known as experimental design, is considered a rigorous design in social science and social policy research (Boruch, 1997; Orr, 1999; Bloom, 2005), seeking to draw a causal inference about treatment effect (Mosteller & Boruch, 2002). In the late 1980s and mid-1990s, RCTs were utilized in several educational interventions to evaluate program effectiveness, helping policymakers to concentrate on important research and distribute funding appropriately (Grissmer, Subotnik, & Orland, 2009). Beginning in 1998, Congress passed several acts advocating rigorous evaluations of educational programs, using either experimental or high-quality quasi-experimental designs (QED). Among these acts, the No Child Left Behind (NCLB) Act, a reauthorization of the Elementary and Secondary Education Act (ESEA) of 1965 that passed in 2001, was the most influential. Further, the movement calling for rigorous evaluation of the impact of educational interventions was pushed forward due to substantial funding provided by the Institute of Education Sciences (IES) (Konstantopoulos, 2008), the research division of the U.S. Department of Education established by the Education Sciences Reform Act of 2002 (ESRA) (Education Sciences Reform Act, 2002).

In 2011, the IES and National Science Foundation (NSF) formed a committee through joint effort aimed at enhancing the quality and knowledge development in science, technology, engineering, and mathematics (STEM) education (IES & NSF, 2013). Two years later, a Common Guidelines for Education Research and Development, simply as the “Common Guidelines” were released by the committee. In the Common Guideline, impact research was highlighted and called for researchers to identify “what works” or what interventions improve student academic outcomes. For impact research, it must “generate reliable estimates of the ability of a fully developed intervention or strategy to achieve its intended outcomes” (IES & NSF, 2013, p14). For an impact study to be warranted, the. This encompasses three types of research: Efficacy, Effectiveness, and Scale-up Research (IES & NSF, 2013). The three types of research are similar in design methods, specified outcome measures, level of statistical power, valid information from intervention for analysis, and counterfactual conditions but they differ in their research purposes, circumstances of interventions delivered, generalizability to populations, and settings (IES & NSF, 2013; Flay, et al., 2005). To begin with, Efficacy Research evaluates whether an intervention has a positive change on student academic outcomes when delivered under ideal conditions, i.e., a homogenous sample of students and schools, with support provided to the teachers or material resources to the classrooms. Next, based on positive evidence from Efficacy research, Effectiveness Research tests whether an intervention has a positive effect on student academic outcomes under routine practice. Similar to Effectiveness Research, Scale-up Research tests whether an intervention has an effect on student academic outcomes, but in such a manner that it can be generalized across populations and settings (IES & NSF, 2013; Flay et al., 2005).

Most recently, experimental design was again given priority in high-quality impact evaluations by the National Center for Education Evaluation and Regional Assistance (NCEE). In its guidelines for evaluation projects (Evaluation Principles and Practices, 2017), NCEE emphasized that an intervention program must demonstrate sound and clear evidence of causal effect inferences in a trial, which also complies with the highest standards of quality for conducting scientifically valid education evaluations, as required by the Education Sciences

Reform Act (ESRA) (Sec.173). Arguably, concern about “what works” has continuously drawn attention in the educational community.

In practice, researchers frequently utilize another type of RCT, called cluster-RCTs in impact research (Borman, Slavin, Cheung, Chamberlain, & Chambers, 2007; Cook, Hunt, & Murphy, 2000). Unlike simple RCTs, where individuals are randomly assigned to treatment and control groups whereas in a cluster-RCT, an intact group (“cluster”) are randomly assigned to different conditions due to the naturally nested structure in school systems, i.e., students nested in classrooms, classrooms nested in schools (Bloom, 2005; Boruch & Foley 2000; Cook, 2005; Hedges, & Hedberg, 2007; Spybrook & Raudenbush, 2009). For example, the National Center for Education Research (NCER) and NCEE within IES have funded 175 RCTs, among which more than 100 studies have adopted cluster-RCT designs in K-12 impact research since 2002 (Spybrook, 2008; Spybrook, 2013). The research areas adopting cluster-RCT designs cover various educational interventions, such as school reforms, curriculum redesigning, or student healthcare programs, in a variety of settings, including pre-K, elementary schools, middle schools, and high schools (“Grantsearch,” 2017). Meanwhile, the demand for rigorous evaluation of the impact of interventions has led to an increased use of RCTs in post-secondary education, especially for community colleges. However, the use of CRT designs for impact research in higher education is a relatively recent phenomenon in review of the IES funded project (“Grantsearch,” 2017).

Whether in educational setting or other contexts, to be a cluster-RCT study, a design must meet two key criteria: (1) the unit of assignment in the study is a cluster, and (2) the data for the analysis is based on individuals within those clusters (WWC, 2013). Historically, cluster-RCTs have been called “cluster randomized trials”, “group-randomized trials” (Murray, 1998) and “place-based randomized trials” (Donner, Brown, & Brasher, 1990; Donner & Klar, 2000). In the early 2000s, the Consolidated Standards of Reporting Trials (CONSORT) statement provided reporting guidelines for RCTs, and the term “cluster randomized trial (CRT)” became the most commonly used for this type of design. Throughout this dissertation, CRTs were used throughout the dissertation in which the “cluster” refer to schools as the randomization takes place at the school level. Although implementation and analysis are equally important in a CRT trial, throughout, the attention is restricted to CRT design by improving power analysis.

Background

Whether in K-12 or higher education, CRTs must be designed with adequate statistical power to produce high quality and rigorous evidence (Hedges & Hedberg, 2007; Spybrook, 2014). An *a priori* power analysis enables researchers to determine the sample size necessary to detect a meaningful effect (Bloom, 1995).

The Importance of Power Analysis

Previous literature has consistently emphasized the importance of conducting a power analysis for CRT studies (Bloom, 2005; Donner & Klar, 2000; Konstantopolous, 2008; Murray, 1998; Raudenbush, 1997; Raudenbush & Liu, 2000; Raudenbush, Spybrook, & Martinez, 2007; Schochet, 2008). One reason is that an underpowered design may result in inconclusive conclusions which can stunt the progress of a field. That is, it will be unclear whether the intervention is not effective or whether the study simply is not powered to detect the treatment effect. Indeed, Song and Herman (2009) addressed the issue that most CRTs in education lack statistical power due to the small number of clusters in the study. However, when more schools or students are recruited than necessary for a study, resources are wasted, which should be avoided because of resource constraints educational communities encountered today (Hedges & Rhoad, 2010; Westine, Spybrook, & Taylor, 2013).

The Importance of Design Parameters

A challenge that researcher faces when planning a CRT is that they must estimate two critical elements in conducting a power analysis: (1) the intra-class correlation coefficient (ICC), and (2) the percent of the variance in the outcome explained by the covariate(s) (R^2) at each level (Hedges & Hedberg, 2007). The ICC and R^2 are also referred to as "design parameters" because of their importance in CRT design stages (Brandon, Harrison, & Lawton, 2013; Jacob, Zhu, & Bloom, 2010). In the context of this dissertation, the ICC is the percent of the total variance in achievement outcomes that is between schools. Take math outcomes, for example. If the ICC for a two-level (students nested in schools) is 0.20, then 20% of the variance in math outcome is between schools and 80% of the variance in math outcome is within schools. The R^2 here is the percent of the variance in the student achievement outcomes that is adjusted by a covariate or set of covariates. Suppose we include a pretest as a predictor, then, if the Level 1(student level) R^2 is 0.178, we can say that 17.8% of variability in math outcomes within schools is explained by pretests. If the Level 2 (school level) R^2 is 0.70, meaning that 70% of the variability in math outcomes between schools is adjusted by pretest scores. For a specified sample sizes, reducing the variance in the outcome will result in the capacity to detect a smaller effect size or the difference between two groups, holding the power constant at 0.80.

Strategies for Estimating Design Parameters

As design parameters are sensitive to different contexts, outcomes, samples, grades, and designs (Bloom, Bos, & Lee, 1999; Bloom, Richburg-Hayes, & Black, 2007; Brandon, Harrison, & Lawton 2013; Hedges & Hedberg, 2014; Jacob, Zhu, & Bloom, 2009; Westine, Spybrook, &

Taylor, 2014; Xu & Nichols, 2010), researchers face uncertainty when estimates are not available for a particular context, which is often the case in higher education. Researchers (Hedges, & Hedberg, 2007; Westine, Spybrook, & Taylor, 2014) recommend three strategies of estimating ICC and R^2: (1) consult the literature for similar studies with reported design parameters, (2) conduct a pilot test with a similar sample to obtain estimates of design parameters, and (3) use large databases to estimate the design parameters. The third method was utilized in this study.

Up to now, scholars have accumulated empirical work around estimating design parameters for impact research in K-12 settings including student achievement outcomes such as reading, math and science. For example, Hedges and Hedberg (2007) produced estimates of ICCs and R^2 using several national datasets that covered name the outcomes and the grades here. Several other scholars estimated ICCs and R^2 based on state-level datasets for K-12 reading, math, and science outcomes (Brandon, Harrison, & Lawton, 2013; Hedges & Hedberg, 2013, 2014; Westine, Spybrook, & Taylor, 2013; Xu & Nichols, 2010; Zhu, Jacob, Bloom, & Xu, 2012). Furthermore, other scholars have provided ICCs and R^2 based on school-district-level datasets (Bloom, Bos, & Lee, 1999; Bloom, Richburg-Hayes, & Black, 2007), while still others have compiled estimated empirical design parameters from past empirical studies or large evaluation studies in K-12 educational interventions (Jacob, Zhu, & Bloom, 2010; Schochet, 2008; Zhu et al., 2012). With those accessible design parameters, researchers can enhance their capacity to plan a CRT design.

Statement of the Problem

While the tremendous progress K-12 has made in documenting design parameters using standardized test outcomes, estimating design parameters for student-learning outcomes in higher education has fallen substantially behind. One potential reason for this delay a lack of agreement on which student learning outcomes are the most important and how to measure these outcomes in higher education (Callen & Finney, 2002; Klein et al., 2005; Shavelson, & Huang, 2003).

Participation in national standardized tests of higher-order cognitive skills, such as critical thinking and writing skills in higher education has increased drastically as accreditation agencies have started to require them in the past decade (Association of American Colleges and Universities and Council for Higher Education Accreditation 2008; Liu, 2010; Steedle, 2012). In addition, universities were called upon to hold themselves accountable by providing evidence of the effectiveness of a college education and student learning outcomes (U.S. Department of Education, 2006). In response to these requirements, two prominent organizations in higher education, the American Association of State Colleges and Universities (AASCU) and Association of Public and Land-Grant Universities (APLU), formerly known as the National Association of State Universities and Lang Grant Colleges (NASULGC), created the Voluntary System of Accountability (VSA) program to measure students' critical thinking, analytical reasoning, and analytical writing abilities using standardized tests (VSA, 2008). Three standardized tests were approved based on their established reliability and validity to measure the VSA-defined core educational outcomes (Liu, 2011). The three standardized tests are: (1) the Collegiate Assessment of Academic Proficiency (CAAP) by American College Testing

(ACT), (2) Collegiate Learning Assessment (CLA) by the Council for Aid to Education (CAE), and (3) Proficiency Profile, formerly known as Measure of Academic Proficiency and Progress (MAPP) by the Educational Testing Service (ETS) (VSA, 2008).

The availability of these standardized tests has led to an increased ability to examine undergraduate students' critical thinking skills and communication skills and to test the effectiveness of interventions designed to improve these outcomes in the past decade. Although intervention for improving cognitive skills maintain momentum, recent studies indicate that a high quality and true experiment designs are few. For example, Behar-Horenstein and Liu (2011) examined forty-one empirical studies in a systematic review of critical thinking skills in higher education. Only three (7%) implemented a true experimental design, fourteen (33%) pre-experimental design, and twenty-five studies (60%) quasi-experimental design. In another systematic review on critical thinking skill instruction interventions in higher education, Tiruneh, Verburgh and Elan (2013) examined thirty-three studies. Four (13%) employed experiment design, and twenty-two (67%) quasi-experiment design, and seven (21%) pre-post design without comparison groups. The drawbacks for such designs are that the studies can be undermined by threats to internal validity such as maturation, dropping out, familiarity with the pre-test, underpowered studies, and so on (Cook & Campbell, 2002; Behar-Horenstein &Liu, 2010; Ennis, 2016). Thus, this study will extend the knowledge of design parameters in K-12 impact studies to higher education in assisting researchers with designing a CRT study in higher education. Although implementation and analysis are equally important in a CRT design (U.S. Education Department, 2003), the scope of this study focuses on improving the power analysis during the design stage.

Purpose of the Study

The purpose of this study is to empirically estimate ICCs and R^2 values for two-level (i.e., students nested in schools) CRTs aimed at evaluating the efficacy of collegiate cognitive skills interventions in higher education. More specifically, the primary outcomes are from a standardized test, the Collegiate Learning Assessment (CLA) test administered between 2005 to 2010 provided by Council Aid to Education (CAE). The CLA test is a holistic and complex standardized test which assesses students' four higher order cognitive skill: critical thinking, analytic reasoning, written communication, and problem solving skills (Arum & Roksa, 2011; Klein, Benjamin, Shavelson, & Bolus, 2007). Particularly, the CLA measures higher order skills that almost all institutions attribute to improving. The primary outcome measures in this study include: Performance Task outcome, Analytical Writing outcome, and total CLA outcome (average of Performance Task outcome and Analytical Writing outcome). The findings from the study are intended to inform the design of two-level trials, in which students are nested within colleges/universities and the unit of randomization occurs at school level.

Research Questions

This current study employed a series of two-level Hierarchical Linear Model (HLM) to calculate design parameters for an important student learning outcome, cognitive skills in higher education. Specifically, the analyses address four central questions as followed:

1. To what extent do the following outcomes vary across schools?:

    a. Performance Task outcome.

    b. Analytical Writing outcome.

    c. The total CLA outcome.

2. To what extent do student-level covariates (i.e., entering academic ability [EAA] score, student demographics, and so on) explain the variance in the three outcomes?

3. To what extent do school-level covariates (i.e., Median SAT) explain the variance in the three outcomes?

4. Given the design parameters estimates in questions 1-3 and effect sizes from the literature, what is the sample size necessary for CRTs that aim to test interventions seeking to improve at colleges/universities?

Contribution to EMR

Previous empirical research on improving the design of CRTs in education is limited to K-12 education within the domain of student academic outcomes in reading, math, and science. Although some scholars published ICCs using CLA tests, there lacks a systematic compilation of empirical design parameters for designing CRTs in higher education. Though we have not seen many CRTs yet to assess the impact of cognitive skill interventions in higher education, we anticipate there will be more as more calls are made for improving these skills for the upcoming workforce. This work is getting out ahead of these calls by starting to build a resource of design parameters in higher education and provide researchers reference values of design parameters when planning a two-level CRT.

Overview of the Dissertation Structure

The overall structure of the study takes the form of five chapters, including this introductory chapter. Chapter II is divided into five sections. The author begins by laying out how the RCT and CRT evolved in K-12 and higher education over the past decades. Then, the author introduces statistical power analysis for 2-level CRTs, which covers MDES approach and empirical design parameters for student academic outcomes in K-12 and higher education. Next, the author talks about the how the covariates were selected based on higher education literature. It ends with a discussion reasonable magnitude of the effect sizes in K-12 and higher education used as empirical benchmark in educational interventions. Chapter III is concerned with the methodology used for this study, which mainly focuses on description about data source and samples, data screening process, outcome measures, covariates, and analytical models. Chapter IV presents the findings of the ICCs, R^2, and MDES aligning with each research question and focus on a discussion of the patterns of ICC and R^2 that might be applied to *a prior* power analysis for the design of a two-level CRT at colleges/universities. Finally, the conclusion in Chapter V gives a summary of the study, as well as areas for further research identified and discussed.

CHAPTER II

LITERATURE REVIEW

To help identify the best practices to improve student academic outcomes in U.S. classrooms and schools, over the past decade, researchers have implemented randomized controlled trials (RCTs) to evaluate educational interventions such as programs, policies, and practices (hereafter referred to as "interventions"). The use of RCTs have grown rapidly since the Institute of Education Sciences (IES) began funding studies with RCT designs in 2002. The movement advocating rigorous research designs such as RCTs was further advanced by the release of the Common Guidelines for Education Research and Development (IES & NSF, 2013), simply referred to as "the Common Guidelines". In particular, the Common Guidelines underscored one type of impact research through rigorous research designs such as RCT approaches.

In K-12 impact research, cluster randomized trials (CRTs)—a special type of RCTs—are becoming more common as a way to test educational interventions that are intended to improve student achievement outcomes. In recent years, higher education has begun favoring CRT designs to test the efficacy of educational interventions such as mathematics interventions, meta-cognitive outcome interventions, among others ("IESgrants", 2017). A key consideration for such a trial is the statistical power to assess an effect of a particular magnitude. However, conducting a power analysis for a CRT is complex. *A priori* estimates of design parameters are required during the planning phase including estimates of the intraclass correlation coefficient (ICCs) and covariate-outcome correlation (R^2). Many scholars (Bloom, Bos, & Lee, 1999; Hedges & Hedberg, 2007; Xu & Nichols, 2010) have documented empirically-estimated design parameters for K-12 reading, math, and science outcomes; many of which are now available in an online compendium to help researchers designing CRTs in K-12 (“variance-almanac-academic-achievement”, 2017). However, higher education lacks such a design parameter compilation to guide researchers with interest of CRT studies. Thus, the purpose of this research is to establish a design parameter repository for two-level CRT in the higher education context. Specifically, the outcome measure of this study was Collegiate Learning Assessment (CLA), one of the most comprehensive standardized tests measuring students’ cognitive growth in higher education (U.S. Department of Education, 2006).

To clarify the role of CRTs in higher education and the importance of relevant and accurate design parameters, seven key aspects from the relevant literature are reviewed here: (1) basic description of RCTs; (2) the evolution of RCTs in K-12 and higher education; (3) CRTs in K-12 and higher education; (4) details related to conducting a statistical power analysis for a CRT, including a discussion of empirical research of ICCs and R^2 in K-12 and higher education; (5) strategies for selecting covariates in higher education; (6) the magnitude of minimum detectable effect size (MDES); and (7) the summary of this chapter.

Description of RCTs

RCTs or experimental designs have been used to evaluate the impact of educational interventions for over 15 years (Spybrook & Raudenbush, 2014). When carefully designed and successfully implemented, RCTs are the most credible research design for establishing causal links between interventions and outcomes (Boruch, 1997; Mosteller & Boruch, 2002; Murnane & Willett, 2010). One important feature of a RCT is the unit of random assignment. The simplest design is one in which individuals are randomly assigned to the treatment or control condition. However, units may also be clusters of people such as classrooms, schools, or districts. A second feature of a RCT design is whether the blocking technique is utilized the study design. Figure 1 presents a person RCT design without blocking. For example, suppose 1,200 low-income freshmen will participate in a RCT to test the efficacy of a first-year experience (FYE) with a mentoring component to improve their GPA and second-year retention rate. Using a person RCT design, the research team randomly assign 600 individuals to treatment (FYE with mentoring components) and another 600 individuals to control group (without mentoring component) (see Figure 1).

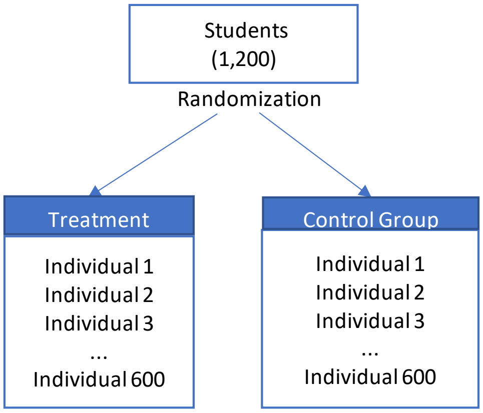


Figure 1. Person RCTs without Blocking

Figure 2 displays a person RCT design with the presence of blocking, also known as a multi-site (or blocked) RCT. The blocking is advantageous in two ways. First, blocking improves the face validity of the experimental study (Spybrook, Bloom, Congdon, Hill, Martinez, & Raudenbush, 2011). Second, blocking is likely to reduce the variability between students within “blocks” and increase the precision of the estimate of the treatment effect. Thus, blocking increases the power of the test for the main effect of treatment (Raudenbush, Martinez, & Spybrook, 2007; Andres & Spybrook, 2009). Use the same example as Figure 1 demonstrates, suppose that 1,200 low-income freshman will participate in the same FYE program. As Figure 2 shows that the research team will arrange students in “blocks” based on orientation sessions to ensure balanced groups of low-income freshmen in both treatment and control groups. The blocking factor, i.e. orientation session in this case, should be one that is strongly associated with outcome variables (Raudenbush, Martinez, & Spybrook, 2007; Andres & Spybrook, 2009). Because of these advantages, multi-site RCTs have also been one of the dominant design for researchers to consider in educational interventions (Spybrook, 2008).

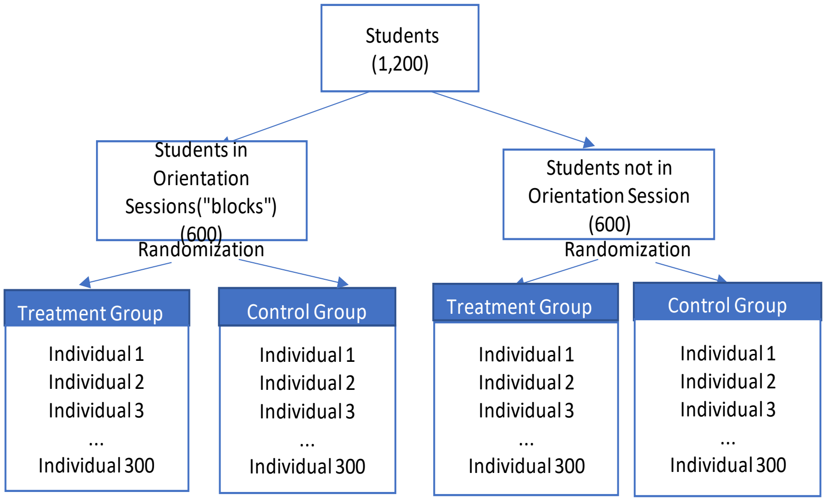


Figure 2. Person RCTs with Blocking

Evolution of RCTs in K-12 Education

In the late 1980-2000 time span, three social and educational interventions with RCT designs were launched to test the effectiveness of intervention programs (Grissmer, 2016; Schultz & Mueller, 2006). One of the initiatives—Tennessee’s Project Student/Teacher Achievement Ratio (STAR)—was the first large, multi-site person RCT in K-12 education, conducted from 1985 to 1989 (Finn & Archilles, 1990). Funded with approximately $12 million from the state legislature and conducted by Tennessee’s State Department of Education, Project

STAR evaluated the effects of class size reduction (CSR) on student achievement. More than 6,000 students in 329 classrooms were involved in the project during its first year, reaching almost 12,000 students over its 4-year duration (Finn & Archilles, 1990; Word, 1990). The key findings indicated STAR's positive impact on student achievement outcomes. Furthermore, the study reduced an achievement gap between minority students and White students. In most cases, minority students gained twice to three times benefits than White students did (Finn & Archilles, 1990; Mosteller, 1995). These positive results provided justification for state and federal CSR programs throughout the nation to improve quality of education (Schanzenbach, 2006). As a pilot study, STAR project was not without flaws. For example, Hanushek (1999) pointed it out that the randomization process was not strictly performed. However, STAR project was an important milestone in education field, demonstrating the usefulness of experimental designs in helping educational community understand the CSR effect on relevant research and policy decision-making (Finn & Archilles, 1990; Mosteller, Light, & Saches, 1996; Sohn, 2016). Mosteller et al. (1996) also recognized the project as one of the great experimental studies in U.S. education history.

In the late 1990s, RCTs were placed on the national education agenda due to several pieces of federal legislation. The first major initiative, the 1998 Obey-Porter legislation, created the Comprehensive School Reform Demonstration (CSRD) program and invested $150 million, calling for scientific evidence in education, which could be provided through the approach high-quality quasi-experiment design (QED) (Borman, Hewes, Rachuba, & Brown, 2002; Borman, 2002; Doherty, 2000). Congress clarified that schools could receive CRSD funding only if they proposed to implement evidence-based educational practices and programs (Borman, Hewes, Rachuba, & Brown, 2002). Because Congress wished to prevent potential harmful effects on children as target subjects due to interventions that had not been proven to be scientifically effective, they passed a second initiative, the Scientifically Based Education Research, Evaluation, and Statistics and Information Act of 2000. Although the Act encountered setbacks, it started establishing standards for both quantitative and qualitative research in education (Boruch & Mosteller, 2002). Consequently, the National Research Council (NRC) formed the committee on Scientific Principles in Education Research, which promoted the use of rigorous methodology in education (Borman, 2002; Mosteller & Boruch, 2002; Lagemann, 1997; Shavelson & Towne, 2002). The continuous demand for high-quality research in education inspired the passing of the 2001 No Child Left behind Act (NCLB), a reauthorization of the Elementary and Secondary Education Act (ESEA) of 1965. Of these three initiatives, NCLB has been the most influential as it stressed evidence-based methods and procedures to enhance the quality of education. Specifically, it recognized experimental designs and QED as acceptable designs for establishing reliable evidence on educational interventions (Borman, 2002).

The movement toward experimental designs was propelled by the passing of the Education Sciences Reform Act of 2002 (ESRA), and by the establishment of the IES, the research division of the U.S. Department of Education (USDOE) (Education Sciences Reform Act, 2002). Particularly, Grover Whitehurst, the first director of IES, was a strong advocate for evidence-based decision-making for improvement of education (Whitehurst, 2002). Beyond that, IES launched a reliable reviewing center, the What Works Clearinghouse (WWC), which evaluated and synthesized research evidence for the effectiveness of educational interventions. Moreover, the WWC set the rating standards for studies under review in the Procedures and Standards Handbook (WWC, Version 4.0). For instance, well-designed and well-implemented RCTs with low attrition are rated as “meets WWC design standards without reservations”. QEDs with equating and no severe design or implementation problems and RCTs with severe design or implementation problems are both rated as “meets WWC design standards with reservations”. Studies that fail to provide sufficient causal evidence for an intervention’s effects are rated as “does not meet WWC design standards”.

In 2003, the USDOE mandated that a “rigorous” study must utilize RCTs or high-quality QEDs (IES, 2003). To reach the goal of improving education quality with strong designs, two IES research centers have invested substantial funding to increase the use of RCT designs: National Center for Education Research (NCER) and National Center for Special Education Research (NCSER). Starting 2002, NCER began supporting rigorous research that identified imperative issues and improved quality of education from preschool (age 3) to adult education (IES Program Overview, 2017). For example, the first NCER funded project is titled *A longitudinal study of the effects of a pre-kindergarten mathematics curriculum on low-income children's mathematical knowledge,* which sought to evaluate the effects of the pre-K mathematics curriculum on low-income children in California and New York in 2002. NCSER began sponsoring rigorous research in 2006 in an effort to build a knowledge base related to infants, toddlers, and students with or at risk for disabilities, from birth through high school (IES Program Overview, 2017). The slight difference between the two centers lies in the target age range for educational interventions. Regarding student academic outcomes, NCER and NCSER both support research on traditional academic outcomes (i.e., reading, writing, math, and science) and social and behavioral competencies that support student success in schools (IES Program Overview, 2017). Additionally, a contract-awarding center, NCEE, is responsible for conducting impact studies, mostly through RCTs to improve student achievement in collaboration with Regional Educational Laboratory Program, WWC, Education Resources Information Center

(ERIC), and National Library of Education (“IESNCEE”, 2017). Together, the NCLB, the ESRA, and WWC have transformed education in the U.S. into an evidence-based research field (Karlet, 2012).

Evolution of RCTs in Higher Education

In higher education, one prominent example of applying RCT is an impact evaluation of the Upward Bound, a federal program designed to increase access to higher education for economically disadvantaged students. Given the fact that higher education was devoid of scientific evidence from previous Upward Bound projects, USDOE commissioned Mathematica Policy Research, Inc. (MPR) to conduct a largest longitudinal impact evaluation of Upward Bound project across nation in 1991, involving 52,000 students participating in 727 these projects (Myers, Olsen, Seftor, Young, & Tuttle, 2004). The study provides empirical evidence on that the Upward Bound project could increase enrollment at four-year colleges and universities as well as had positive impact on high school math credits earned by participants but not on other outcomes (Myers, Olsen, Seftor, Young, & Tuttle, 2004). In the same year, the USDOE started the Experimental Sites Initiative (ESI) and Congress authorized the ESI under section 487A (b) of the Higher Education Act (HEA) of 1965 as amended (“experimentalsites”, 2017). The Initiative allows institutions to test and evaluate policy changes on a small scale to advise future higher education legislation (“experimentalsite”, 2017).

However, in a recent report *Putting the experimental back in the Experimental Sites Initiative* (2018*)*, authors Mccann, Laitinen and Feldman asserted that USDOE launched 30 “experiments” through the Initiative but only two programs were designed with credible evaluations. On the other hand, because evidence-based educational practices and programs were not prioritized by the administration or mandated by Congress, the higher education community lacked clear guidance and standards on how to conduct RCT studies to inform broader decision-making for relevant policies (Mccann, Laitinen, & Feldman, 2018). Indeed, Ross, Morrison, and Lowther (2005) had earlier raised concern about methodological issues in experimental studies in higher education such as low internal and external validities of experiment designs, lacking theoretical framework and analytical techniques to strengthen experimental designs. Consequently, relatively few experimental studies have been published in peer-reviewed journals compared with that in K-12 body of relevant literature, which might stunt the progress of RCTs in higher education.

The rise of RCTs to test the impact of educational interventions in higher education mainly stemmed from community colleges in the early 2000s. Historically, community colleges have been viewed as the best way for low-income individuals to achieve a higher education and improve their prospects for the labor market (Goldberg & Finkelstein, 2002; Gordon, Young, & Kalianov, 2001). Although considerable interventions to increase completion rate have been tried over the decades (Goldberg & Finkelstein, 2002; Gordon, Young, & Kalianov, 2001; Tinto, 1997, 1998; Zhao & Kuh, 2004), few studies were well-designed to support causal inferences on those interventions (Bettinger & Baker, 2011; Evans, Kearney, Perry, & Sullivan, 2017; Schultz & Mueller, 2006). In response to these issues, MDRC (formally known as "Manpower Demonstration Research Corporation"), a pioneer in advancing rigorous evaluation to measure the effects of social and educational policy initiatives, introduced the “Open Door Demonstrations” project in 2003 (Scrivener & Coghlan, 2011). One of the large-scale experiment studies was to evaluate the effectiveness of the Kingsborough Community College’s (KCC) learning community program conducted in 2003-2005 (Visher, Weiss, Weissman, Rudd, & Wathington, 2012; Weiss, Mayer, Cullinan, Ratledge, Sommo, & Diamond, 2014). Because of the solid evidence the KCC learning community program produced, learning community interventions were later scaled up to other community colleges (Weiss, Mayer, Cullinan, Ratledge, Sommo, & Diamond, 2014). Furthermore, funded by a consortium of foundations, the U.S. Department of Labor, and the USDOE, MDRC implemented several impact evaluation programs with RCT designs striving for improving student academic success such as various forms of student services programs (Hock, 2010; Scrivener &Weiss, 2009) and financial incentives intervention programs (Brock & Richburg-Hayes, 2006; Richburg-Hayes et al., 2009).

Among all impact evaluation programs performed by MDRC, the City University of New York's (CUNY) Accelerated Study in Associates Program (ASAP) was the exemplar of how application of RCTs could be used to test the efficacy of community college completion interventions. Findings of the project revealed that the ASAP study almost doubled graduation rates from 22 percent to 40 percent after three years (Scrivener, Weiss, Ratledge, Rudd, Sommo, & Fresques, 2015). Because of its success, ASAP also received great attention in the higher education field. Further, to help understand whether ASAP could be implemented to other settings and populations to boost completion rates, Great Lakes Higher Education Guaranty Corporation funded a replication demonstration of the ASAP model (2014-2019) in three community colleges in Ohio, which was led by MDRC and CUNY ("mdrcproject", 2017).

With growing concern over college enrollment and completion rates has heightened interest, Congress passed the Higher Education Opportunities Act (HEOA), which required USDOE implemented best "promising practices" in research to increase high school students' access to higher education and increase completion rate (Higher Education Opportunity Act, 2008). In response to the call, higher education saw an increase of impact researcher, which mostly were funded by the Postsecondary and Adult Education (PSAE) grants under NCER research center. To date, NCER has invested over $95 million to support 59 research projects., Twenty-nine out of 59 (49%) are impact research. Especially, 15 out of 29 projects adopted RCT designs including the ASAP project mentioned above (“IESprograms”, 2017). Because of all the above-mentioned efforts, a fair amount of literature on experimental studies started to emerge in higher education.

As the on-going requirement for evidence-based studies, RCTs did not only serve as a robust design to establish causal inference of educational interventions in community colleges but also in four-year colleges/universities. Some recent empirical studies include performance-based studies (Binder, Krause, Miller& Cerna, 2015), developmental summer bridge programs (Barnett, Bork, Mayer, Pretlow, Wathington, Weissman, & Teres, 2015), and mentoring programs (Bettinger & Baker, 2011; Steeg, Elk, & Webbink, 2014), just to name a few. In particular, an experimental evaluation of student mentor program conducted by Bettinger and Baker (2011) demonstrated the positive effect of intensive student mentoring on increasing two- and four-year persistence and degree completion rates, this study was later frequently cited by applicants to the Strengthening Institutions Program (SIP) and the First in the World (FITW) grants offered by the USDOE. Collectively, the RCT studies for community colleges and four-year colleges/universities have added to the growing body of literature on experimental studies and provided methodological guidance to educational interventions in higher education.

Today, recognition of the importance of RCTs in higher education research continues to gain momentum. In 2017, the Center for Research on Undergraduate Education (CRUE) Symposium invited scholars from across the nation to share their innovative rigorous evaluations of interventions in higher education; topics included but not limited to college access, admissions, STEM achievement, and student success initiatives. Subsequently, in a Higher

Education Act reauthorization bill introduced in December 2017 by the chairwoman of the House Education and the Workforce Committee, emphasis on reviving the mission of ESI by mandating rigorously designing and evaluating experiments to inform decision-making in higher education were address clearly (Fredman, 2018). Hopefully, with ongoing input from all aspects in higher education, the revitalization of RCTs in rigorous evaluation will be pushed forward.

Description of CRTs

In addition to the person RCT, the two most widely used experimental designs in education research are variation of RCTs: CRT and multi-site (blocked) CRT designs. When cluster of individuals (e.g, students nested in classrooms) are the unit of random assignment in an experimental design, it is referred to as a CRT design. If a CRT design includes blocking, it is a multi-site (blocked) CRT design. Both designs assume a clustered sampling design but differing in terms of how random assignment is made to treatment or control conditions (Hedge & Rohads, 2010). For this dissertation, the author focused on a two-level CRT.

A CRT design, also known as a “hierarchical design” (HD), is when “clusters” at Level 2 (e.g., classrooms or schools) are randomly assigned to the treatment or control conditions. Figure 3 presents a 2-level CRT design (e.g, 200 individuals nested in each school). As can be seen that the randomization occurs at Level-2 (school level). Six schools as “clusters” are randomly assigned to treatment and control conditions as the arrow directed. Hence, all individuals within a given cluster (school) receive the same treatment. CRT designs have won favor of researchers in educational interventions for several reasons. First, a CRT design can potentially reduce contamination by physically separating individuals receiving different treatments (Raudenbush, Martinez, & Spybrook, 2007). Contamination occurs when interaction between individuals in different treatment conditions causes some individuals to receive features of a treatment to which they were not assigned. For instance, if an experimental study is designed to increase gateway science courses retention rates by introducing a redesigned curriculum intervention delivered to some students but not others within a given university, it is possible that the students receiving the intervention will also have their learning experience shared with peers not receiving the interventions.

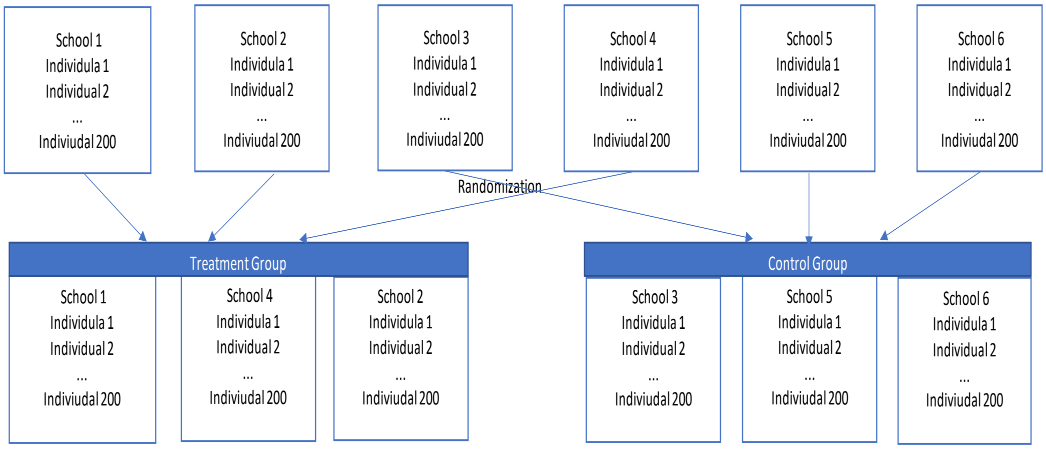


Figure 3. Two-level CRT Design with Students Nested in Schools

Second, because many educational interventions such as whole-school initiatives are deployed at the entire school environment or classroom level, a 2-level CRT often makes the most sense to test these types of interventions (Bloom 2005; Boruch & Foley, 2000; Cook, 2005). For example, in a Connected Chemistry Curriculum (CCC) intervention program for high school chemistry, a CRT is well-suited because the interventions are delivered at the classroom context to investigate how students engaged in chemistry and the measure the effects of student learning outcomes (Mike, Superfine, &Yin, 2017). In addition to the scientific value provided by CRTs, implementation of this design poses few logistical, ethical, and administrative challenges while maintaining the integrity of the study in practice (Bloom, 2005; Raudenbush, 1997). As

Cook and Payne (2002) addressed that administrators and district officials tend to participate in an experiment when entire schools or districts receive the treatment.

In summary, researchers can choose appropriate designs based on the pros and cons of CRT designs into consideration. With careful attention to the issues addressed, researchers can avoid methodological pitfalls and use these approaches successfully. As the need for high-quality education evaluations was expressed by the NCEE (Evaluation Principles and Practices, 2017), it is anticipated that CRT designs will continue to thrive in impact research.

CRTs in K-12 Education

Since 2002, IES has funded a new generation of intervention studies that adopted CRTs, beginning with K-12 intervention programs. In 2003, five CRTs were funded by IES ("IESgrant", 2017). Today, 111 out of 266 impact research studies are through the approach of CRT designs (both simple CRT and blocked CRT), thanks to funding provided by the NCER and NCSER centers ("IESgrant", 2017). The research covers a variety of topics including education programs, practices, and policies in reading and writing, mathematics, and science education; teacher quality; cognition and student learning; and high school reform, among others ("NCERGrant", 2017). While NCEE has initiated approximately 34 impact studies, with 16 studies implementing CRTs covering a variety of research topics such as early literacy, mathematics, teacher quality, special education, and English language learning, among others ("NCEEGrant", 2017).

CRTs in Higher Education

In contrast, the progress of CRT applied in higher education has fallen behind compared to K-12 research. The first CRT that evaluated the impact of student achievement outcomes in post-secondary education was probably the Beacon Mentoring Program at South Texas Colleges, performed by MDRC in spring 2008 to evaluate the impact of student academic outcomes in mentored classes. In the Mentoring Program, the unit of randomization took place at 83 sections of developmental (remedial) math or college algebra either to receive a mentor to be part of the control group (Quint, 2011). It was not until 2016, there appeared the first CRT design in academic domain funded by NCER titled *Supporting strategic writers: effects of an innovative developmental writing program on writing and reading outcomes,* a multisite CRT with random assignment of instructors within college to treatment and control groups. So far, 5 out of 22 experimental studies were employed CRT designs and all were supported by both NCER and NCEE centers. The research topics include but were not limited to: student access to, persistence in, progress through, and completion of postsecondary education ("NCER programs", 2017).

As the momentum for promoting innovative solutions and evidence in post-secondary education persists, one important effort to foster innovative ideas that help keep college increase quality and improve educational outcomes is the previously mentioned FITW program funded by the USDOE since 2014. The first FITW grant project using a CRT design was awarded to Spelman College's metacognitive training program, which incorporated new teaching and learning strategies to test the effectiveness of student metacognitive training in both classroom and peer-tutoring settings ("FITWgrant", 2017). Also, in her recent testimony submitted to the Senate Committee on Health, Education, Labor, and Pensions, Dr. Lashawn Richburg-Hayes (2015) suggested that much like in K-12, higher education needed to increase the use of CRTs to test the impact of interventions and build a base of knowledge in a topic area, i.e., in financial aid reforms to help low-income students achieve academic success. This innovative idea was also amended by Congress under Section 487A (b) of the Higher Education Act of 1965, to recruit colleges and evaluate the interventions through randomized trials. Furthermore, to help researchers in planning CRTs, IES has made CRT design tools and technical assistance publicly available for evaluations of interventions funded by the Investing in Innovation and FITW programs ("IES Evaluation TA", 2017). Given the recent call for rigorous evidence stemming from higher education initiatives, it is hoped that more CRT designs will be welcome in higher education.

Statistical Power Analysis for Two-level CRTs

The power of statistical test is the probability that it will yield a statistically significant result, which is set at .80 in social sciences by convention (Cohen, 1988). A statistical power analysis is a method of determining the probability that a proposed research design will detect an expected effect of a treatment (Hedge & Rhoads, 2010). The importance of conducting a power analysis for CRT studies cannot be emphasized enough. As the large-scale impact studies involves million-dollar investment, a power analysis before a study can reduce a potential waste of resources by collecting and analyzing data from a sample larger than necessary (Konstantopoulos, 2009; Westin, Spybrook, &Taylor, 2013). Whereas a study with insufficient power may result in a wrong conclusion that the intervention does not have a significant impact when it actually does (Shadish, Cook, & Campbell 2002). In addition, it is common to include a power analysis in grant proposals to justify the sufficient sample sizes in the proposed study that can generate expected effect (IES, 2009; NSF, 2009; Scheier & Dewey, 2007).

There are two primary approaches to conducting a power analysis: power determination approach and effect size approach. The first one aims to calculate the power for a given sample size and determined effect size. Whereas the second one aims to calculate the minimum detectable effect size (MDES) for a given sample size and determined power depending on specific contexts (Spybrook et al., 2011). In this study, MDES approach was utilized for the power analysis.

MDES Approach for Power Analysis

Originally, the impact estimate of a CRT design is measured in its original unit called minimum detectable effect (MDE)—the smallest true effect that is likely being found to be statistically significant (Bloom, 1995). In order to compare impacts across different outcome variables, the MDE is usually reported in units of standard deviations, known as the MDES (Bloom, 1995). In a two-level CRT, the MDES is a function of five components: (1) the statistical power(1-β), (2) the alpha level (a), (3) the number of students per school (n) and the number of schools (J), (4) effect size (δ), and (5) ICCs and R^2 (Spybrook, et al., 2011). By convention, the statistical power is set at 0.80 for a two-tail test; and the alpha level (a) at 0.05 in education. However, ICC and R^2 as two key design parameters play consequential roles in *a prior* power analysis for a CRT study as they have to be estimated before a study.

To help CRT researchers acquire essential skills for power analysis, many scholars have made invaluable contributions to in this area for different types of CRTs (Bloom, Bos, & Lee, 1999; Donner & Klar, 2000; Murray, 1998; Hedges & Hedberg, 2007; Raudenbush & Liu, 2000, 2001; Raudenbush, Spybrook, Liu, & Congdon, 2006; Snijder & Bosker, 1993, 1999; Spybrook et al., 2011; Spybrook, 2014). For example, Raudenbush (1997) identified two important findings for two-level CRT power analysis: (1) the number of clusters has more influence on the power than the number of individuals per cluster, and (2) the higher the ICCs, the lower the power for a given number of clusters. Findings from other studies also were consistent with Raudenbush’s (Bloom, Bos, Lee, 1995; Bloom, 2005; Hedges & Hedberg, 2007; Hill, Bloom, Black, & Lipsey, 2007). In essence, the smaller of MDES a CRT study can detect, the more precise of the study is, holding statistical power constant at .80. Thus, MDES is also regarded as a gauge to assess the precision of a CRT study (Spybrook & Raudenbush, 2009).

Empirical Design Parameters for Student Academic Outcomes

In practice, the biggest challenge a power analysis of a two-level CRT is to obtain ICCs and R^2 values, which are unknown and must be estimated before a study because these values are specific to grades, subjects, and school settings (Bloom, Bos, & Lee, 1999; Bloom et al., 2007; Bloom et al., 2008; Brandon, Harrison, & Lawton 2013; Hedges & Hedberg, 2014; Jacob, Zhu, & Bloom, 2009; Westine, Spybrook, & Taylor, 2014; Xu & Nichols, 2010). Previous empirical literature discussed design parameters for student academic domains in K-12 educational research. In the following section, the author briefly summarizes work relevant to two-level design parameters.

Design Parameters in K-12. Scholars have extensively investigated ICCs for math, reading, and science outcomes across different grades and school districts (Bloom, Bos, & Lee, 1999; Bloom, Richburg-Hayes & Black, 2007; Hedges & Hedberg, 2007; Konstantopoulos, 2009; Schochet, 2008; Westine, Spybrook, &Taylor, 2013). In addition, scholars discussed the important function of covariates. Especially, covariates with strong predicative power can reduce the sample sizes needed to achieve adequate power, holding all other parameters constant. In turn, the covariate can reduce the cost of a study under consideration (Bloom, Richburg-Hayes, & Black 2007; Hedges & Hedberg 2007; Raudenbush, Martinez, & Spybrook, 2007).

Table 1 presents empirically estimated magnitude of ICCs and R^2. As can been seen that Bloom, Bos, and Lee (1999) were pioneers in documenting ICCs for math and reading in Grades 3 to 6. They also found that student and school level covariates had strong explanatory power for improving the precision of the study. Since then, many efforts have been made in investigating how different design parameters can have effect on different types of CRT designs and it has

Table 1

Magnitude of Empirical ICCs and R^2 for Two-Level CRT in K-12

Source	Student academic outcome	Grade	Range of ICCs	Range of R^2
Bloom, Bos, and Lee (1999)	Math and Reading	3-6	0.14 -0.21	..
Bloom, Richburg-Hayes, and Black (2007)	Math and Reading	3,5, 8,10	0.13-0.29	0.07-0.52 (R^2_{L1}) 0.18-0.89 (R^2_{L2})

Hedge and Hedberg (2007)	Math and Reading	K-12	0.15-0.25	0.22-0.52(R^2_{L1})
				0.30-0.73(R^2_{L2})

Schochet (2008)	Math and Reading	1-6	0.10-0.20	0-0.50 [a]
Konstantopoulos (2009)	Math and Reading	K-5	0.10-0.25	..

Westine, Spybrook and Taylor (2013)	Science	4,5,8,10,11	0.17-0.31	0.07-0.13(R^2_{L1})
				0.53-0.87(R^2_{L2})

Hedge and Hedberg (2014)	Math and Reading	1 to 11	0.02- 0.43	0.57-0.64(R^2_{L1})
				0.80-0.87(R^2_{L2})

Note. [a]Schochet (2008) used assumed group level variance 0.0-0.50(R^2_{L2}). ".." indicates that R^2 values are not specified.

concluded that design parameters vary across samples, outcomes, or grades (Bloom, Richburg-Hayes & Black, 2007; Hedges & Hedberg, 2007; Konstantopoulos, 2009; Schochet, 2008; Westine, Spybrook, &Taylor, 2013). Particularly, Hedges and Hedberg (2007) implied that choosing covariates from variables that were correlated with the outcomes without being influenced by the treatment including: student/school level pretests, demographic variables (i.e., age, gender, race/ethnicity, socio-economic status [SES]), and indicators of school challenges such as English language learner's status. Of these covariates, the most effective covariates are one-year lagged student/school level pretest (Bloom et al., 2007). Moreover, the authors claimed that depending on the context, student level and school level covariate can explain about 50% to 80% or more variance at each level. Beyond that, a website called the Variance Almanac of Academic Achievement ("Web VA") provides access to a comprehensive compendium of design parameters for reading and math by the Center for Advancing Research and Communication (Hedge & Hedberg, 2007). These design parameters were gleaned from various datasets ranging from kindergarten to12th grade across the nation (Hedge & Hedberg, 2007).

In summary, the recommended ICC magnitude ranges between 0.15 and 0.25 based on U.S. datasets on school achievement to help researchers justify the design of CRTs (Bloom, Bos, & Lee, 1999; Bloom, Richburg-Hayes, & Black, 2007; Hedges & Hedberg, 2007; Schochet, 2008). Regarding the magnitude of outcome-covariate variance (R^2) at each level, it can take the value as much as 0.50 to 0.80 (Hedges & Hedberg, 2007; Bloom, Richburg-Hayes, 2007).

Design Parameters in Higher Education. In contrast, few empirical estimates of ICCs and R^2 exist in the body of higher education. This is likely a result of several challenges facing the higher education community. First, there is a lack of guidelines for reporting and interpreting ICCs in higher education, which makes them less practically meaningful (Niehaus, Campbell, & Innkeals, 2013; Dedrick et al., 2009). Second, higher education often relies on graduation rates, retention, endowment level, student/faculty ratio, etc. to measure institutional effectiveness (Klein, Kuh, Chun, Hamilton, & Shavelson, 2008; Gates et al., 2001). There lacks a universal definition or measurement (operationalization) of graduation and retention across universities (Van Stolk et al., 2007). Further, these outcomes are often not continuous in nature and hence traditional ICC calculations may not be relevant. Of the empirical work that exist in higher education, Niehaus, Campbell and Innkeals (2013) investigated the magnitude of ICCs which varied from 0.001 and 0.33 including both two-, and three-level HLM models. Still, quite a few scholars reported ICC but didn't calculate how much the proportion variance explained by including Level 1 covariates (R^2_{L1}) or Level 2 covariates (R^2_{L2}) for two-level HLM in higher education. Table 2 presents the empirical ICCs and R^2 for two-level CRT in higher education. Note that some of the R^2_{L1} and R^2_{L2} were not directly reported in the articles cited but manually

Table 2
Magnitude of Empirical ICCs and R^2 for Two-Level CRT in Higher Education

Source	Student Outcomes	Outcome Measures	Range of ICCs	Range of R^2
Kim (2001)	Students' social desire To influence social condition	CSS	0.23	..
Hu and Kuh (2003)	Student learning productivity	CSEQ	0.085	..
Kinzie, Thomas, Palmer, Umbach, and Kuh (2007)	Satisfaction of education experience	NSSE	0.02-0.12	..

McCormick, Kuh, Pike, and Chen (2009)	Cognitive skills outcome
	Non-cognitive gains
	Academic challenges
	Active and collaborative learning
	Student-faculty interaction enriching education experience
	Supportive campus environment
		NSSE	0.045-0.177	0.004-0.428(R^2_{L1})
				0.006-0.660(R^2_{L2})

Liu (2011)	Cognitive skills outcome	PP	0.14-0.17	0.16-0.47 (R^2_{L1}) 0.44-0.68 (R^2_{L2})
Steedle (2012)	Cognitive skills outcome	CLA	0.19-0.26	0.03-0.06 (R^2_{L}) 0.87-0.95(R^2_{L2})

Note. College Student Survey=CSS; College Student Experience Questionnaire=CSEQ; National Student Survey of Engagement=NSSE; Proficiency Profile=PP.

calculated by the author based on within school and between school variances without or without covariates using Formula 16 and 17. Overall, the magnitude of ICCs for two-level HLM gathered in this study ranges from 0.02 to 0.26, which is somehow close to that in K-12; R^2_{L1} from 0.004 to 0.47; and R^2_{L2} from 0.006 to 0.95, respectively (Kim, 2001; Hu & Kuh, 2003; Kinzie, Thomas, Palmer, Umbach, & Kuh, 2007; Liu, 2011; McCormick, Kuh, Pike, & Chen, 2009; Steedle, 2012).

Even though all these design parameters reported made great contribution to the literature in higher education, there still needs a comprehensive and compilation of design parameters for researchers to justify the sample sizes needed for a CRT study, especially for one of the most important student learning outcomes, cognitive skill outcomes in higher education. Thus, this study will fill the void by starting to build one for cognitive skill outcomes in higher education.

Selecting Covariates in Higher Education

Unlike in K-12 CRT studies, selecting covariates is more challenging due to various definitions, measures, interventions to improve those higher-order skills in higher education. For example, Pascarella and Terenzini (1991, 2005) conducted extensive research on factors associated with critical thinking gains in higher education. While Arum and Roksa's (2011) study used the CLA test to investigate factors related to cognitive skill gains, a test measures not only critical thinking skill but also written communication and complex reasoning skills. Despite the difference, three aspects related to cognitive skill outcomes in higher education were considered: students' pre-college characteristics, students' experience at colleges/universities, and institutional characteristics (Arum & Roksa, 2011; Astin, 1993; Berger & Milem, 2000; Pascarella &Terenzini, 1991; 2005). The Conceptual Framework (see Figure 4) shows relationships between factors and the cognitive skill outcomes. It also guides the flow of selecting covariates related to the cognitive skill outcome in this study.

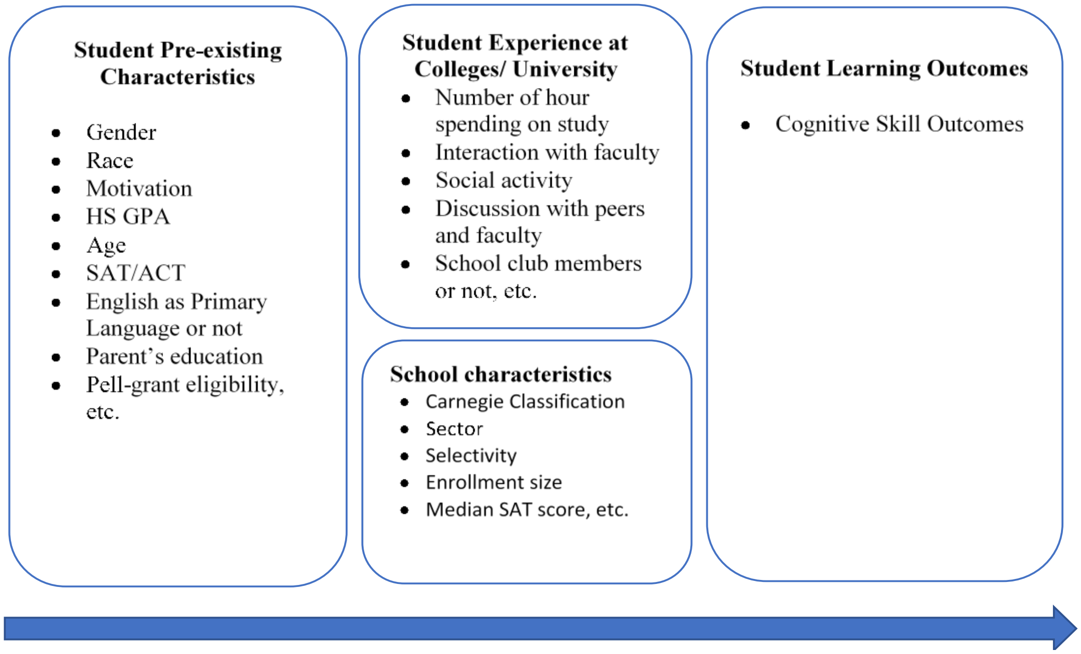

Student Pre-existing Characteristics
Gender
Race
Motivation
HS GPA
Age
SAT/ACT
English as Primary Language or not
Parent's education
Pell-grant eligibility, etc.
Student Experience at Colleges/ University
Number of hour spending on study
Interaction with faculty
Social activity
Discussion with peers and faculty
School club members or not, etc.
School characteristics
Carnegie Classification
Sector
Selectivity
Enrollment size
Median SAT score, etc.
Student Learning Outcomes
Cognitive Skill Outcomes

Figure 4. Conceptual Framework for Covariates Related to Cognitive Skill Outcome in Higher Education

Student Pre-existing Characteristics

Pre-existing characteristics appearing in the literature include gender, race/ethnicity, ability, motivation, parental support, high school GPA (HS GPA) (Astin, 1993; Atkinson & Geiser, 2009; Berger & Milem, 2000; Kuh, 2009; Weidman, 1989; Pascarella, 1985), and Society Economic SES (Cunha & Miller, 2014), among others. For example, HS GPA and SAT/ACT scores can serve as proxies for pretests because they are usually not available in higher education literature (ACT, 2009; Kobrin et al., 2008; Rothstein, 2004; WWC, 2014). Especially, HS GPA has been recognized strong predictor for first-year college GPA, which accounts for 30% of the variance in first-year college GPA (Atkinson, 2001; Kobrin, et al., 2008). Moreover, proxies for SES in higher education can include indicators such as free lunch, parents' education, Free Application for Federal Student Aid (FAFSA), expected family contribution, family income, Pell grant eligibility, and first-generation college student status (WWC, 2014).

Students' Experience at Colleges/Universities

Higher education saw an emergence of various models that examined students' experience associated with cognitive growth at schools. For instance, Pascarella (1985) developed a college impact model that established a direct relationship between institutional characteristics to the college environment. Weidman (1989) suggested accounting for academic characteristics (e.g., mission, selectivity, and major) and social characteristics such as family SES to test the effect of schools on students' cognitive growth. Astin (1993) focused on factors associated with individual, structural, organization related to students' cognitive skills improvement.

Previous studies have indicated that academic effort and engagement such as time spent studying and reading can improve students' cognitive skills (Carini & Kuh 2003; Kuh et al., 1991; Terenzini et al., 1995). Astin (1993) found that the number of hours spent studying was positively related to all self-reported increases in cognitive ability. In terms of social integration, student interactions with peers and faculty enhance students' capacity for solving complex cognitive tasks (Astin, 1993; Chickering, 1969; Kuh, 1995; Terenzini & Pascarella, 1980; Pace, 1990; Volkwein, King, & Terenzini,1986). Arum, Roksa, and Velez (2008) identified factors associated with improvement in critical thinking skills using CLA Performance Task. In general, students' perception of faculty expectations is positively related to critical thinking skills. In addition, the authors found that academic preparation such as HS GPA attributed to better cognitive skills improvement whereas involvement in fraternities or sororities had negatively related to cognitive skill gains. Aside from it, the author found that students' majoring in math, science, social science, and the humanities were advantageous in critical thinking skills than students in other fields of study.

School Characteristics

Scholars had studies on institutional characteristics (e.g., selectivity, mission, and sector, etc.) have effect on students' cognitive skills development with inconsistent results. Weidman (1989) suggests considering academic characteristics (e.g., mission, selectivity, and major) and social characteristics such as family SES to evaluate the effect of college on students' cognitive growth. Drawn on CLA data, Klein et al. (2008) examined the effect of institutional characteristics on student critical thinking skills improvement. Their findings suggested that institutional characteristics accounted for 10% of the variance in the senior mean CLA score. However, Steedle (2011) pointed it out that when controlling for entering academic ability (EAA) scores at student level using a two-level HLM, most institutional factors (e.g., sector, selectivity, full-time retention rate, etc.) were inconsequential on CLA outcome. In addition, they found students' motivation only accounted for about 3% to 7% of the variance in school-level outcomes (Klein, et al., 2007). Moreover, critical thinking skills measured by the Performance Task was correlated with National Survey of Student Engagement (NSSE) but not with strong magnitude (Carini, Kuh, & Klein, 2006). Different from these scholars' findings, Pascarella and Terenzini (2005) found that college selectivity had a negligible effect on cognitive skills development based on ACT's Collegiate Assessment of Academic Proficiency (CAAP) test.

Given that various factors attributed to students' outcomes in higher education, O'Connell and Reed (2011) suggested that choosing covariates should depend on research questions, theory, consultations with relevant stakeholders rather than solely on the results of a statistical test (O'Connell & Reed, 2011). Thus, the covariates chosen were built on previously cited literature and availability of covariates in the dataset. More detailed illustration about the rationale for each covariate can be found in the “Methodology” section.

The Magnitude of the Effect Size

As noted above, the goal of power analysis in this study is to determine MDES—the smallest true effect a treatment can detect in standard deviation units that is likely being found to be statistically significant (Bloom, 1995). This section discusses the well-established MDES for student academic outcomes in K-12 education and prior empirical effect size for cognitive skill in higher education.

Empirical Effect Sizes in K-12 Education

Empirical effect size for student achievement outcomes has been well-established in K-12 educational interventions. For instance, an effect size of 0.20 implies an impact or treatment effect equal to one fifth of student level standard deviation of the outcome across all students from all schools in a CRT study (Bloom et al., 2005). One accepted magnitude of effect size in social science is Cohen’*s d* (1969), which defines 0.20, 0.50, and 0.80 be considered small, medium, and large, respectively. Whereas in educational field, scholars provided more empirical guidelines for specific domains and target population accordingly. For example, scholars in NCES (1977) recommended that the empirical benchmark of mean effect sizes for high school students’ annual growth were: 0.17 for reading and 0.26 for math nationwide. Moreover, Lipsey (1990) examined 186 meta-analyses of 6,700 studies and revealed that the distribution of effect sizes was almost identical between non-educational and education research. That is, the small effects ranged from 0.00 to 0.32; medium effects ranged from 0.33 to 0.55; and large effects ranged from 0.56 to 1.20.

In the following years, scholars continuously investigated the magnitude of effect size in educational interventions based on robust experimental designs. For instance, Kane (2004) provided reference for nationwide reading and math improvement and suggested an average of 0.25 standard deviation (*SDs*) was appropriate. Moreover, Hill, Bloom, Rebeck-Black, and Lipsey (2007) specified a range of effect sizes from 0.20 to 0.30 which was regarded as plausible in educational interventions. For policy-decision making, Bloom, Richburg-Hayes, and Black (2005) argued that effect sizes between 0.10 and 0.20 for student achievement might be considered. Most importantly, the authors insisted on that effect sizes should refer to empirical benchmarks that are relevant to the intervention, target population, and outcome measure. Most importantly, they recommended three types of effect size benchmarks: (1) expectations for growth or change in the absence of an intervention, (2) policy-relevant gaps compared to existing differences among subgroups of students or schools, and (3) impact findings from previous research on similar grade levels, interventions, and outcomes.

Empirical Effect Size for Cognitive Skills in Higher Education

Empirical studies on cognitive skills have been studied intensively in higher education (Astin, 1993; Carini & Kuh, 2003; Ishiyama, 2002; Kim & Sax, 2009, 2011; Kitchener, Wood, & Jensen, 2000; Pascarella & Terenzini, 2005; Twale & Sanders, 1999; Volkwein, Valle, Parmely, Blose, & Zhou, 2000; Whitmire, 1998). The common practice in higher education is to report the expected effect size for growth without an intervention. For example, Pascarella and Terenzini (2005)'s study mainly focused on one component of cognitive skills, critical thinking skill gains using ACT's CAAP test. The scholars found students' critical thinking skill gains varied by their stay in universities. On average, the effect size was from .55 and .65 *SD*s in four-year but less than 1*SD* in synthesis of several studies. In the Wabash National Study, the mean effect size on

CAAP critical thinking skill was .11 *SD*s during the first year and .44 *SDs* over four years (Pascarella, Blaich, Martin, & Hanson, 2011). For CLA expected change in effect size, Arum, Roksa, and Velez's longitudinal study (2008) found an average increase of .18 *SD*s on the CLA Performance Task during the first two years and a four-year of .47 *SD*. However, they claimed that the effect sizes were not sufficient for measuring students' growth in critical thinking skills.

Another effect size benchmarking is to compare with similar cognitive skill interventions using similar definitions, measures, intervention, treatment intensity, samples, and designs. For example, Ortiz's (2007) reported gains of .12 *SDs* per semester for nonphilosophy students. Niu, Behar-Horenstein, and Garven (2013) reported an effect size of .195 *SD* for 12 weeks based on 31 empirical studies focusing on instruction interventions on college students' critical thinking gains. In a most recent study, Huber et al. (2016) examined 71 studies and estimate the overall effect of college students' critical thinking skills is at 0.59 *SD* when compared nursing with non-nursing students. Different from Arum and Roksa's viewpoint, Huber et al. (2016) regarded 0.59 *SD*s sufficient improvement of critical thinking skills during colleges. Collectively, the effect size for critical thinking skill interventions falls into a range from 0.11 to 0.59 *SD*s in the body of literature in higher education.

However, cautioned about the use and interpretation of these effect sizes should be taken seriously since they were context specific (Bloom, et al., 2005; Pascarella, et al., 2011). First, when using empirical cognitive skill effect size as benchmarking, one should note that the cognitive skills often interchangeably with "critical thinking" in literature but can be assessed with different measures. Second, because most of the effect size yielded from observational studies rather than robust experimental designs, one should be careful of the meaningful magnitude of the effect size. Finally, the effect sizes were calculated based on different samples of students and institutions, measures of critical thinking skills, and statistical analysis methods (Pascarella, et al., 2011). Thus, the effect sizes on cognitive skills outcome should refer to empirical benchmarks that are relevant to specific definitions, interventions, target population or subpopulations, and outcome measures (Bloom, at el., 2005).

Summary

As CRTs continue to be driving force behind K-12 impact research, they have also come to influence higher education research. Although some scholars report empirically estimated values of ICCs for Performance Task outcomes, there still needs a systematic collection of ICCs and R^2 for CRT design purposes in higher education. To test the efficacy of programs intended for increasing cognitive skills in higher education, it is necessary to begin to develop empirical estimates of design parameters so that CRTs can be planned with adequate statistical power. Building on the valuable experience of designing CRTs trials accumulated in K-12 impact research, this study is a first step towards developing this repository of design parameters for a two-level CRT trial in higher education.

CHAPTER III

METHODOLOGY

This chapter begins by reviewing the purpose and research questions of the study. This is followed by describing data source and analytical sample, data screening, outcome measures, and covariates considered in the models. Also, a series of two-level HLM models with students nested within schools are presented. The formula of calculating ICCs, R^2, and MDES (with and without covariates) are also presented. This section ends with a summary of the chapter.

Review of Purpose and Research Questions

The purpose of this study is to empirically estimate values of ICCs and R^2 using two-level HLM models which aims at evaluating the efficacy of cognitive skill interventions in higher education. More specifically, the primary outcomes are the total CLA outcome, the Performance Task outcome, and the Analytical Writing Task outcome as measured by the CLA test. First, the variance for each outcome was decomposed across students and colleges/universities for each outcome. Second, the percent of the variance was estimated for each outcome explained by covariates at each level under the same two-level data structure. Third, power calculations were demonstrated based on the results of the findings from Questions 1-3, which are documented to be used for planning two-level trials on improving cognitive skill interventions in higher education. The research addresses the following four questions:

1. To what extent are the following outcomes clustered in colleges/universities:
   A. The total CLA outcome?
   B. Performance Task outcome?

C. Analytical Writing outcome?

2. To what extent do student-level covariates (i.e., EAA demographic variables) explain variance in the three outcomes?

3. To what extent do school- level covariates (i.e., Carnegie classification, median SAT, sector, etc.) explain the variance in the three outcomes?

4. Given the design parameters estimates in Questions 1-3 and effect sizes from the literature, what are the sample sizes necessary for a two-level CRT trials which aims to test the efficacy of cognitive skill interventions in colleges/universities?

Data Sources and Samples

The data consists of two sources for student-level and school-level variables. The students' CLA test scores and other administrative data were provided by CAE, which included seven variables in this study: the total CLA outcome, the Performance Task outcome, the Analytical Writing outcome, EAA, English as primary languages, gender, race/ethnicity, and age. The data collection included three phases as presented in Table 3. In Phase1, data were collected when freshman took the test in the fall of 2005 or the fall of 2006. This data was denoted as Sample A, which included 37 schools and 9,827 students. In Phase 2, data were collected in the spring of 2007 or the spring of 2008. Students in Phase 2 were those who were completed sophomore courses. Not all students in Phase 1 participated in Phase 2. Students who had outcome scores in both Phase 1 and Phase 2 were tracked, which were denoted as Sample B, consisting of 27 schools and 2,422 students. In the same manner, in Phase 3, data were collected from senior students in the spring of 2009 or the spring of 2010. Students who had outcomes scores in all phases were denoted as Sample C, which included 22 schools and 1,064 students.

By tracking the same students over time in Sample B and C, it is expected to look at the changes in design parameters across years for a stable sample.

There were 9, 827 students at the onset of the CLA longitudinal study, then shrunk drastically to 2,422 and 1,064 in the follow-up years (see Table 3). Since the CLA dataset in this study was a secondary data source, it was hard to surmise any possible reasons for such a drastic change in sample sizes. One way to understand this phenomenon is to consult the similar studies conducted by other researchers. For example, in the CLA Lumina Longitudinal Study, Klein, Steedle, and Kugelmas (2010) stated that the problem of high drop-out rate of participating schools rose due to difficulties in recruiting and retaining schools, which was usually the case most of the longitudinal studies encountered in higher education.

Table 3
Student and School Sample Sizes by Phase in this Study

	Phase 1 (fall 2005 or fall 2006)		Phase 2 (spring 2007 or spring 2008)		Phase 3 (spring 2009 or spring 2010)	
Sample	n	J	n	J	n	J
Sample A	9,827	37	..	..	..	..
Sample B	2,422	27	2,422	27	..	..
Sample C	1,064	22	1,064	22	1,064	22

Note. n=number of students; J=number of schools; "..”=no participants in specific time. Sample A were freshmen; Sample B were students who completed sophomore courses; Sample C were seniors.

The second data source for institutional characteristics comes from the Integrated Postsecondary Education Data System (IPEDS), which was organized by the National Center for Education Statistics (NCES). The CLA datasets in this study had already linked student level data to school level data from IPDES with identifiers being masked. Generally, the institutional samples in IPEDS are representative of four-year, not-for-profit colleges and universities in the United States. Six school-level variables were considered in this study: median SAT, Carnegie Classification, sector, mean student-related expenditures per full time equivalent (FTE) student, the percent of freshmen receiving Pell grants, and enrollment size of an institution.

Data Screening

Prior to the final analysis, SPSS 25 software was utilized to clean data and prepare datasets to ensure the data quality. Taken student-level and school-level together, there were 13 variables were included for data screening.

Accuracy, Normality, Outlier, Linearity, and Multiclonality

Descriptive and inferential statistics were first performed to portray and analyze the data on the thirteen variables in a flat table format. Then the accuracy of data entry, tests for normality, outliers, and linearity focused on the three outcomes were performed referring to Tabachnick and Fidell’s (2001) data cleaning checklist. All outliers were removed from the dataset once identified. Aside from it, a bivariate correlation analysis indicated that the percent of Pell grant recipients and the median SAT were strongly correlated (*r*=0.835). Collinearity diagnostics also indicated one of the condition indices among multiples variables was 44.113 (greater than the threshold 30), suggesting presence of collinearity. Moreover, multicollinearity diagnostic test was performed by examining the indicator variance inflation factor (VIF) among the variables. The results show that there exists multicollinearity among variables because value of VIF for the percent of freshmen receiving Pell grant variable and median SAT variables both exceeded 4, which can be problematic as it increases the variance of the regression coefficients, making them unstable (Montgomery, 2001; O’Brien, 2007). According to the extant literature in

CLA studies, median SAT variable was the most effective covariate in explaining variance in students' CLA outcomes compared with percent of freshmen receiving Pell grants. Especially, when considering median SAT in two-level HLM models, many institutional variables (e.g., sector, enrollment size, Carnegie Classification, etc.) were not significant factor in explaining the variance in the outcomes. Thus, the percent of freshmen receiving Pell grants was removed from the analytical models.

Dealing with Missing Values

The missing values in each variable cannot be underestimated as they pose threats to the internal validity (e.g., statistical power) and external validity (McKnight et al., 2007; Robert & Karen, 2005). Particularly, variables with missing values above 5% can affect parameter estimates and lead to errors in inference and interpretation of the analysis results (Tabachnick & Fidell, 2001). Table 4 indicates missing values in for student level variables in Sample A, B, and C across Phase 1-3. Overall, missing values for student level variables ranged as low as 0.65% in the Performance Task outcome to as high as 20.30 % in the total CLA outcomes. And no missing values in school level variables were identified.

Rather than ignoring them, the first step is to detect the missingness pattern, be it Missing Completely at Random (MCAR), Missing at Random (MAR), or Missing not at random (MNAR) (Little & Rubin, 1988). The Little's MCAR test was conducted and the results showed that the missing pattern were not MCAR (χ^2 (1, 9827) =2144.702, ρ= .000). This is not surprising since MCAR is rare in reality (Little & Rubin, 2002; Schafer & Graham, 2002). Since MCAR test is out, MAR can be inferred, but missingness is predictable from the variables (Tabahnick & Fidell, 2012). The next step was to see whether the data met the assumption of the MAR pattern.

Table 4

Percent of Missing Values for Each Variable

Variable	N	Number of missing values	Percent of missing values
Sample A (Phase 1)			
Entering Academic Ability	9,711	116	1.19%
Gender	9,356	471	5.03%
Race/Ethnicity	9,356	471	5.03%
Age	9,165	662	7.22%
Total CLA outcome	8,169	1,658	20.30%
Performance Task outcome	9,764	63	0.65%
Analytical Writing outcome	8,212	1,615	19.67%
Sample B (Phase 1)			
Entering Academic Ability	2,409	13	0.54%
Gender	2,422	0	0.00%
Race/Ethnicity	2,422	0	0.00%
Age	2,325	97	4.00%
Total CLA outcome	2,133	298	11.93%
Performance Task outcome	2,412	10	0.41%
Analytical Writing outcome	2,412	580	11.56%
Sample B (Phase 2)			
Entering Academic Ability	2,409	13	0.54%
Gender	2,422	0	0.00%
Race/Ethnicity	2,422	0	0.00%
Age	2,421	1	0.04%
Total CLA outcome	2,293	129	5.33%
Performance Task outcome	2,410	12	0.50%
Analytical Writing outcome	2,302	120	4.95%

Table 4—Continued

Variable	N	Number of missing values	Percent of missing values
Sample C (Phase 1)			
Entering Academic Ability			
Gender	1,064	0	0.00%
Race/Ethnicity	1,064	0	0.00%
Age	1,028	36	3.38%
Total CLA outcome	986	78	7.33%
Performance Task outcome	1,060	4	0.38%
Analytical Writing outcome	990	74	6.95%
Sample C (Phase 2)			
Entering Academic Ability	1,059	5	0.24%
Gender	1,064	0	0.00%
Race/Ethnicity	1,064	0	0.00%
Age	1,063	1	0.05%
Total CLA outcome	1,059	5	0.24%
Performance Task outcome	1,061	3	0.15%
Analytical Writing outcome	1,026	38	1.84%
Sample C (Phase 3)			
Entering Academic Ability	1,059	5	0.47%
Gender	1,064	0	0.00%
Race/Ethnicity	1,064	0	0.00%
Age	1064	4	0.38%
Total CLA outcome	1,043	21	1.97%
Performance Task outcome	1,061	3	0.28%
Analytical Writing outcome	1,046	18	1.69%

Note. N=sample size of participants;

As Tabahnick and Fidell (2012) addressed since MAR is an untestable assumption, the validity of the analysis results depends on the strength of this assumption over the observed variables. To detect the missing value pattern, a *t* tests was conducted to check whether there existed associations between missingness for outcome variables and the values of other variables in the datasets. For instance, the missingness for Performance Task outcome was significantly associated with other variables and the variables used for imputation. This finding further supports the assumption prevalent in higher education that MAR pattern is most feasible because researchers usually have information about individual participates' mobility, perceptions about school processes, and student academic outcomes from previous studies (Schafer & Graham, 2002; Cox, McIntosh, Reason, & Terenzini, 2014). Based on the assumption of MAR pattern for missing values, a sensitivity analysis was conducted to test the robustness of ICCs based on primary analysis utilizing two different technique: listwise deletion and multilevel multiple imputation (MI). HLM 7.02 software was utilized to handle missing data using listwise deletion function built in the software. That is, all data for a case that had one or more missing values were removed (Peugh & Enders, 2004). For incomplete multilevel data, some scholars suggested that the imputation model take the multilevel structure into account to ensure valid statistical inferences in the final multilevel analyses (Black, Harel, & McCoach, 2011; Graham, 2012; Van Buuren, 2011). The R pan package (Schafer & Yucel, 2002; Schafer & Zhao, 2014) and the R package jomo (Quartagno & Carpenter, 2016) can be used for multilevel multiple imputation. The R jomo software package was considered categorical variables existed in the datasets (Grund, Lüdtke, & Robitzsch, 2016). Table 5 presents the ICCs generated on ANOVA model data structure after taking missing values into consideration. There was little variation in ICCs for all three outcomes regardless of which technique was utilized. Therefore, the decision was made to adopt the listwise deletion method for all models as it is a widely used method.

Table 5

Sensitivity Analysis for ICCs Based on the ANOVA Model

	Multilevel Multiple Imputation			Listwise Deletion		
Outcomes	σ^2	τ	ICCs	σ^2	τ	ICCs
CLA	19802.04	7584.67	0.277	19797.24	7590.38	0.277
Performance Task	15345.61	7033.97	0.314	15326.11	6924.57	0.311
Analytical Writing	27164.51	7274.13	0.211	26437.55	6748.36	0.203

Note. σ^2 =student level variance; τ=school-level variance; J (school sample size)=8,061, n(student sample size)=37 after removing missing values by technique of listwise deletion.

To summarize, the data screening procedure improved data quality and the cleaned-up data were ready for final statistical analysis. Collectively, a total of thirteen variables were retained in the final analysis: total CLA outcome, Performance Task outcome, Analytical Writing outcome, EAA, race/ethnicity, gender, English as primary language, age, Carnegie Classification, sector, median SAT, mean student-related expenditure per FTE student, and enrollment size of the institution.

Outcome Measures

Differing from the traditional format of standardized tests utilized in K-12, most of the tests given at colleges/universities use a multiple choice or true/false format to test student learning outcomes. The CLA test assesses four higher-order skills in college students: critical thinking, analytic reasoning, written communication, and problem-solving. Specifically, three outcome measures (scaled measures treated as continuous variables) were examined in each phase for a given sample: (1) a Performance Task (PT) measures students’ analytical reasoning and evaluation, problem solving, writing effectiveness and writing mechanics by asking students' to draft a letter, memo, or similar document; (2) an Analytic Writing Task (AW) measures students' skills in articulating complex ideas, examining claims and evidence, and supporting ideas with relevant reasons and examples, cohesive discussion, and using standard English by Make-an-Argument and Critique-an-Argument questions, (3) the total CLA (on a scale of 400 to 1600), which is the average of Performance Task and Analytical Writing scores (Assessment, C. L., 2008; CAE, 2009).

Scaling Process

Given that PTs and AWTs scores are of different difficulty levels, raw total scores from the different tasks are converted to a common scale of measurement using a linear transformation to make comparisons across tasks possible (Assessment, C. L., 2008; CAE, 2009).

Reliability

Because the CLA protocol relies upon a matrix sampling approach, the CAE provides each school with guidance on strategies for achieving a representative sample. Specifically, CAE recommends that schools test at least 100 students, or 25-50% of the population size for each class level to ensure reliability of the test results (CAE, 2009). Further, Klein et al. (2008) identified that the reliabilities for the analytic writing tasks reached 0.82 at student level and 0.91 at school level means, respectively. The reliability for performance task were 0.84 at student level and 0.92 at school level means, respectively. However, Klein et al. addressed that one caveat was that these scores were highly reliable when the unit of analysis is the institution and data are aggregated at the institutional level due to the matrix sampling approach applied in the CLA test (Assessment, C. L., 2008).

Validity

In 2008, CAE conducted the Test Validity Study in concerted efforts with ACT and ETS to investigate the construct validity of these three assessments (Klein, Liu, et al., 2009). Overall, the results from the study indicated that critical thinking measured by Performance Task (CLA) was correlated with critical thinking skills measured by equivalent tests conducted by ACT and ETS range from .73 to .83 (Klein, Liu, et al., 2009).

Covariates

Covariates selected into the analytical models were following the conceptual framework as displayed in Figure 4. However, it is important to note that due to the availability of data in the dataset, some variables would like to be considered, such as HS GPA, motivation and academic activities (e.g., number of hour spending on study, interaction with faculty, etc.) were unavailable. Table 6 shows original variables and coding as well as recoded indicator variables considering in the analytical models.

Level-1 Covariates

Level 1 covariates includes five variables in the analytical models: EAA, gender, English as a primary language, race, and age. EAA scores were converted SAT scores (Math + Verbal) or ACT Composite scores on a common scale produced by CAE. Thus, EAA scores were used as proxies for pretests controlling for pre-existing differences in academic abilities. The models accounted for English as primary language by maintaining original coding (1=yes, 0=no). Gender was recoded into indicator variable (1=male, 0=female). Similarly, race/ethnicity was dummy-coded with White (non-Hispanic) as the reference group takes the value of “0”; each non-reference group was recoded from original numeric variables to take the value of “1”, indicating the presence of the effect.

Table 6

Original Variable and Recoded Categorical Variables in This Study

Original Variable	Recoded Variable
Level 1	
English as primary language: 1=yes, 0=no	1=yes, 0=no
Gender: 1=male, 2=female	1=male, 0=female
Race/Ethnicity: 1=Black, non-Hispanic, 2=American Indian/Alaska Native, 3=Asian /Pacific Islander, 4=Hispanic, 5=White, non-Hispanic, 6=Other	1=Minority (Black, non-Hispanic, American Indian, Alaska Native, Asian, Pacific Islander, Hispanic, Other), 0=White, non-Hispanic, reference group
Level 2	
Carnegie Classification: 1=Baccalaureate Colleges, 2=Master's Colleges/Universities, 3=Doctorate-Granting Institution	1=Baccalaureate Colleges or Master's Colleges/Universities, 0=Doctorate-Granting Institution, reference group
Sector: 1=Public, 2=Private	1= Public, 0=Private
Mean student-related expenditures per FTE student	1=$5,000 or less 2=between $5,001 and $10,000 3=between $10,001 and $15,000 4=between $15,001 and $20,000 5=between $20,001 and $25,000 6=between $25,001 and $30,000 7=between $30,001 and $35,000 8=more than $35,000
Enrollment size of institution 1=Small [up to 3,000], 2= Midsized [3,001-10,000], 3=Large [10,001 or more])	1=Small (up to 3,000) or Midsize 0=Large as reference group

Note. CAE provided the original codes for each variable in the CLA dataset; FTE=Full-time equivalent.

Level-2 Covariates

Level-2 covariates included five variables in the analytical models: Carnegie Classification, median SAT score, sector, mean student-related expenditures per FTE student, and enrollment size of institution. Mean student-related expenditures per FTE student was treated as continuous variables. In terms of the median SAT score (i.e., school-level pretest), it was approximately estimated by averaging the 25th and 75th percentile scores (Kugelmass & Ready, 2010). For schools which only reported ACT scores, these scores can be converted to SAT scores by referring to the concordance table on the College Board website. Other categorical variables were recoded into indicator variables such as Carnegie Classification, sector well as enrollment size of institution.

Analytical Models

Seven two-level HLMs were employed to estimate the design parameters empirically. For illustrative purposes, Table 7 displays the descriptors for each model aligning with the research question, student- school-level covariates as well as ICCs, R^2_{L1} and R^2_{L2}. Note that due to restricted space, Table 7 presents the original variable labels for demonstration purpose. Actual design parameter calculation was based on recoded indicator variables displayed in Table 6.

Research Question 1: Unconditional Model 1

Question 1 investigated to what extent the outcomes varied across schools in each outcome. To address a fully unconditional model (without covariates), an ANOVA model with random effect at Level 2 generated restricted maximum likelihood (RML) estimates of variance components, which provided information about the variation in the student academic outcomes within and between universities. ICCs were calculated based on the within and between variances generated by the unconditional model. In the same manner, ICCs were calculated for

Table 7

Covariate Description in the Models

Questions	Models	Student-level Covariates	School-level Covariates	Design Parameters
Question 1	Model 1	..	..	ICCs
Question 2 (student-level covariate)	Model 2	EAA	Mean EAA	$R^2_{L1;}$ R^2_{L2}
	Model 3	Gender Race English spoken as primary language	Mean gender Mean race Mean English as primary language	$R^2_{L1;}$ R^2_{L2}
	Model 4	EAA Gender Race English as primary language	Mean EAA Mean gender Mean race Mean English as primary language	$R^2_{L1;}$ R^2_{L2}
Question 3 (school-level covariates)	Model 5	..	Median SAT	..; R^2_{L2}
	Model 6	..	Carnegie classification Sector Enrollment Size Mean student-related expenditure	..; R^2_{L2}
	Model 7	..	Median SAT Carnegie classification Sector Enrollment Size Mean student-related expenditure	..; R^2_{L2}

Note. In the statistical models, the covariates at the student-level will be aggregated to the school-level. ".." indicates no covariates.

written communication outcomes and the total CLA outcomes. Referring to Raudenbush and Bryk's (2002) notation, the model is presented as below:

Model 1 Level 1: $Y_{ij} = \beta_{0j} + r_{ij}$ $r_{ij} \sim N(0, \sigma^2)$ (1)

Where:

Y_{ij} is the student academic outcome for student i at university j
β_{0j} is the mean student academic outcome for university j
r_{ij} is the random error associated with student i at university j, var $(r_{ij}) = \sigma^2$.

Level 2: $\beta_{0j} = \gamma_{00} + u_{0j}$ $u_{0j} \sim N(0, \tau_{00})$ (2)

Where:

γ_{00} is the grand mean achievement outcomes across universities
u_{0j} is the random error associated with universities means, var $(u_{0j}) = \tau_{00}$.

Model 1 is based on three assumptions. First, it is assumed that the outcomes follow a normal distribution with school-specific means (β_{0j}) and a common variance (σ^2) within all schools. The existence of this common variance constitutes the homogeneity of variance assumption. Second, it is assumed that the school means differ based on a normal distribution with an overall mean γ_{00} and variance τ_{00}. The third assumption is based on that there is no correlation between the residuals at Level 1 and those at Level 2. All else being equal, the larger σ^2 is, the greater the individual variability in student academic outcomes within universities. Similarly, all else being equal, a larger τ_{00} would indicate a large amount of variability between universities in terms of an average student academic outcome.

ICC Calculation Formula. For each sample and outcome domain, analyses were based on Model 1(ANOVA model) to calculated ICC by utilizing HLM 7.02 software. The ICC assesses how strongly the clusters (schools) contribute to the dependency in the data with a range of between 0 and 1. If the ICC takes the value of "1", the observations within each cluster are the same; and if ICC takes the value of "0", all the observations are statistically independent, indicating that Level 2 has no influence on Level 1. To set the variance on standardized scale, the

ICC formula for Sample A at Phase 1, Sample B at Phase 1-2, and Sample at Phase 1-3 are as followed:

$$\rho(\mathrm{ICC}) = \frac{\tau_{00_{\mathrm{unconditional}}}}{\tau_{00_{\mathrm{unconditional}}} + \sigma^2{}_{\mathrm{unconditional}}} \tag{3}$$

$$= \frac{\text{between school variance}}{\text{between school variance+within school variance}}$$

$$= \frac{\text{between school variance}}{\text{total school variance}}$$

Where τ_{00} is between school variance, σ^2 is within school variance, and ρ (ICC) is the percent of the total school variance in student achievement outcomes that is between schools.

To measure the uncertainty associated with these estimates, the standard error can be calculated using the approximation presented in Donner and Koval (1983):

$$\mathrm{SE}\ (\rho) = \sqrt{\frac{2(1-\rho)^2[1+(n-1])\rho]^2}{n(n-1)J}} \tag{4}$$

Where n is the total number of participants per school, J is the total number of schools, and ρ is the ICC.

Research Question 2: Conditional Models 2-4

Question 2 was set up to investigate the extent to which the student-level covariates (i.e., EAA and demographic variables) explained the variability in each of the three outcomes. As such, the author calculated the proportion of variance explained at Level 1(R^2_{L1}) and Level 2 (R^2_{L2}). R^2_{L1} was calculated as a function of the percentage of the variance by accounting for covariate(s) at Level 1 over unconditional models. In the same manner, R^2_{L2} was calculated as a function of the percentage of the variance by accounting for covariate (s) at Level 2 over unconditional models.

In this study, the author explored three scenarios accounting to examine the explanatory power of covariates: (1) pretests (EAA), (2) a composite of demographics, and (3) both pretests (EAA) and demographics. Note that in all cases the individual level covariates are included at level 1 and aggregated up to Level 2. Aside from it, the covariates in all models were not centered as the analytical results showed no difference on the estimates of the variance components in the models, which further supported the viewpoint maintained by Spybrook, Westine, and Taylor (2013). In the following section, the author started with the pretest (EAA) covariate model.

Model 2: The Pretest (EAA) Model. This model includes EAA at Level 1 and it was aggregated at Level 2.

Model 2 Level 1: $\mathrm{Y_{ij}} = \beta_{0j} + \beta_{1j}\mathrm{EAA_{ij}} + \mathrm{r_{ij}}$ (5)

Where

$\mathrm{Y_{ij}}$ is the student academic outcome for student i at university j
$\mathrm{EAA_{ij}}$ is the pretest covariate for student i at university j
β_{0j} is the mean outcome for university j
β_{1j} is the effect of $\mathrm{EAA_{ij}}$ for university j
$\mathrm{r_{ij}}$ is the random error associated with student i at university j, controlling for $\mathrm{EAA_{ij}}$
var ($\mathrm{r_{ij}}$)= $\sigma^2_{|X}$,but this is now a conditional or residual variance.

Level 2: $\beta_{0j} = \gamma_{00} + \gamma_{01}(\text{Mean EAA})_{\mathrm{j}} + \mathrm{u_{0j}}$ (6)
$\beta_{1j} = \gamma_{10}$

Where:

γ_{00} is the adjusted grand mean student achievement outcomes across all universities
γ_{01} is the effect of Mean EAA across all universities
γ_{10} is the average Mean EAA-achievement outcome regression slope across all the universities
$\mathrm{u_{0j}}$ is the random error associated with university means var ($\mathrm{u_{0j}}$)= τ_{00}, controlling for Mean EAA, but this is now a conditional or residual variance $\tau_{00|\mathrm{W}}$.

Model 3: Demographics Model. This model considers the case in which no pretests are available but the demographic variables available can serve as covariates.

Model 3 Level 1: $\mathrm{Y_{ij}} = \beta_{0j} + \beta_{1j}\mathrm{Gender_{ij}} + \beta_{2j}\mathrm{Race_{ij}} + \beta_{3j}\mathrm{English} + \mathrm{r_{ij}}$ (7)

Where:

$\mathrm{Y_{ij}}$ is the student achievement outcome for student i at university j
$\mathrm{Gender_{ij}}$ is the indicator for the sex of student i at university j (1=male, 0=female)
$\mathrm{Race_{ij}}$ is the indicator for the race of student i at university j (1=Minority, 0=White as reference group)
$\mathrm{English_{ij}}$ is the indicator for English as primary language or not for student *i* at university (1=yes, 0=no)
β_{0j} is the mean achievement at university j
$\beta_{1j}$is the "gender" gap at university j, i.e., the mean difference between the achievement of male and female students
β_{2j} is the "minority" gap at university j, i.e., the mean difference between the achievement of white and minority students
β_{3j} is the differentiating effect of English as primary language students vs. English language learners at university j
$\mathrm{r_{ij}}$ is the random error associated with student i at university j, controlling for gender, race, and English as primary language students, var ($\mathrm{r_{ij}}$)= $\sigma^2_{|X}$, but this is now a conditional or residual variance.

Level 2: $\beta_{0j} = \gamma_{00} + \gamma_{01}(\text{Mean Gender})_j + \gamma_{02}(\text{Mean Race})_j + \gamma_{03}(\text{Mean English})_j + u_{0j}$ (8)

$\beta_{1j} = \gamma_{10}$
$\beta_{2j} = \gamma_{20}$
$\beta_{3j} = \gamma_{30}$

Where:

γ_{00} is the adjusted grand mean student achievement outcomes across all universities
γ_{01} is the effect of Mean Gender (1=male, 0=female) for all universities
γ_{02} is the effect of Mean Race (1=Minority, 0=White, non-Hispanic, reference group) for all universities
γ_{03} is the effect of Mean English as primary language for all universities
γ_{10} is the Mean Gender-achievement outcome slope for all universities
γ_{20} is the Mean Race-achievement outcome slope for all universities
γ_{30} is the Mean English as primary language (1=yes, 0=no) students-achievement outcome slope for all universities
u_{0j} is the random error associated with the university means, controlling for the Mean

Gender, Mean Race, and Mean English variables, var $(u_{0j})=\tau_{00}$, but this is now a conditional or residual variance $\tau_{00|w}$.

Model 4: The Pretest (EAA) and Demographics Model. In this model, both pretests and demographic covariates are considered.

Model 4 Level 1: $Y_{ij} = \beta_{0j} + \beta_{1j}EAA_{ij} + \beta_{2j}Gender_{ij} + \beta_{3j}Race_{ij} + \beta_{4j}English_{ij} + r_{ij}$ (9)

Where:

Y_{ij} is the student academic outcome for student i at university j
EAA_{ij} is the continuous covariate for student i at university j
$Gender_{ij}$ is the indicator for the sex of student i at university j
$Race_{ij}$ is the indicator for the race of student i at university j
$English_{ij}$ is the indicator for English as primary or not for student i at university j
β_{0j} is the mean achievement at university j
β_{1j} is the average change in achievement outcome for a -unit increase in EAA_{ij} for university j
$\beta_{2j}$is the "gender" gap at university j, i.e., the mean difference between the achievement of male and female students
$\beta_{2j}$is the "gender" gap at university j, i.e., the mean difference between the achievement of male and female students
β_{3j} is the "minority" gap at university j, i.e., the mean difference between the achievement of white and minority students
β_{4j} is the differentiating effect of English as primary language students vs. English language learners at university j
r_{ij} is the random error associated with student i at university j, controlling for EAA, Gender, Race, and English as primary language students, var $(r_{ij})= \sigma^2_{|X}$, but this is now a conditional or residual variance.

Level 2: $\beta_{0j} = \gamma_{00} + \gamma_{01}(MeanEAA)_j + \gamma_{02}(Mean\ Gender)_j + \gamma_{03}(Mean\ Race)_j + \gamma_{04}(Mean\ English)_j + u_{0j}$ (10)

$\beta_{1j} = \gamma_{10}$
$\beta_{2j} = \gamma_{20}$
$\beta_{3j} = \gamma_{30}$
$\beta_{4j} = \gamma_{40}$

Where:

γ_{00} is the adjusted grand mean student achievement outcomes across all universities
γ_{01} is the effect of Mean EAA, i.e., the average increase or decrease in mean outcomes, $\beta_{0j,}$ for students

γ_{02} is the effect of Mean Gender, i.e., the difference in mean outcomes, $\beta_{0j,}$ for male students compared with female students
γ_{03} is the effect of Mean Race, average difference in mean outcomes β_{0j} for white students compared with minority students
γ_{04} is the effect of Mean English as primary language, i.e., the average difference in the mean outcomes for English as primary students compared with English leaner students
γ_{10} is the effect of Mean Gender-achievement outcome slope for all universities
γ_{20} is the effect of Mean Race-achievement outcome slope for all universities
γ_{30} is the effect of Mean English (as primary students)-achievement outcome slope for all universities
γ_{40} is the effect of Mean English as primary students-achievement outcome slope for all universities
u_{0j} is the random error associated with the university means, controlling for Mean EAA, Mean Race, Mean Gender, and Mean English as primary language, var $(u_{0j)}=\tau_{00,}$ but this is now a conditional or residual variance $\tau_{00|W}$.

Research Question 3: Conditional Models 5-7

Question 3 was used to investigate to what extent the school-level covariates explain the variability in each of the three outcomes. Models 5-7 were used to address the question.

Model 5: The Median SAT Model. Model 5 takes the scenario of no administrative data into account. In other cases, researchers have no access to this type of information or it is expensive to obtain. Including institutional characteristic variables is a feasible way to overcome this issue as those data are publicly accessible on the IPEDS website. The author started by including median SAT in the school level, which is a covariate that is likely to have strong explanatory power. As the equation demonstrates below, no Level 1 covariates are included in the models with the Level 2 covariate, i.e., median SAT scores.

Model 5 Level 1: $Y_{ij} = \beta_{0j} + r_{ij}$ (11)

Where:
Y_{ij} is the student academic outcome for student i at university j
β_{0j} is the mean achievement at university j
r_{ij} is the random error associated with student i at university j, var $(r_{ij})= \sigma^2$

Level 2: $\beta_{0j} = \gamma_{00} + \gamma_{01}(\text{Median SAT})_j + u_{0j}$ (12)

Where:

γ_{00} is the adjusted grand mean student achievement outcomes across all universities
γ_{01} is the effect of Median SAT for all universities
u_{0j} is the random error associated with the university mean, controlling for Median SAT, var $(u_{0j)}=\tau_{00,}$ but this is now a conditional or residual variance $\tau_{00|W}$.

Model 6: Institutional Characteristics Model. In this scenario, school-level pretests (median SAT scores) was considered in the model.

Model 6 Level 1: $Y_{ij} = \beta_{0j} + r_{ij}$ (13)

Where:

Y_{ij} is the student academic outcome for student i at university j
β_{0j} is the mean achievement at university j
r_{ij} is the random error associated with student i at university j, var $(r_{ij})= \sigma^2$.

Level 2: $\beta_{0j} = \gamma_{00} + \gamma_{01}(\text{Carnegie})_j + \gamma_{02}(\text{Sector})_j + \gamma_{03}(\text{Expenditure})_j + \gamma_{04}(\text{Size})_j + u_{0j}$

Where:

γ_{00} is the adjusted grand mean student achievement outcomes across all universities
γ_{01} is the effect of Carnegie Classification (1=Baccalaureate Colleges or Master's Colleges/Universities, 0=Doctorate-Granting Institution as reference group for all universities)
γ_{02} is the effect of Sector(1= Public, 0=Private) universities
γ_{03} is the effect of mean student related expenditure per FTE student for all universities
γ_{04} is the effect of enrollment size for universities (1=Small (up to 3,000) or Midsize 0=Large as reference group)
u_{0j} is the random error associated with the university means, controlling for Carnegie, Sector, Expenditure, and Size, var $(u_{0j})=\tau_{00,}$ but this is now a conditional or residual variance $\tau_{00|w}$.

Model 7: Median SAT and Institutional Characteristics Model. The last scenario considered both the median SAT and institutional characteristic covariates at Level 2.

Model 7 Level 1: $Y_{ij} = \beta_{0j} + r_{ij}$ (14)

Where:

Y_{ij} is the student academic outcome for student i at university j
β_{0j} is the mean achievement at university j
r_{ij} is the random error associated with student i at university j, var (r_{ij})= σ^2.

Level 2: $$\beta_{0j} = \gamma_{00} + \gamma_{01}(\text{Median SAT})_j + \gamma_{02}(\text{Carnegie})_j + \gamma_{03}(\text{Sector})_j + \gamma_{04}(\text{Expenditure})_j + \gamma_{05}(\text{Size})_j + u_{0j} \quad (15)$$

Where:

γ_{00} is the adjusted grand mean student achievement outcomes across all universities
γ_{01} is the effect of Median SAT for all universities
γ_{02} is the effect of Carnegie Classification for all universities
γ_{03} is the effect of Sector, i.e., private and non-profit universities vs. public universities
γ_{04} is the effect of mean student-related expenditure per FTE student for all universities
γ_{05} is the effect of enrollment size for all universities
u_{0j} is the random error associated with the university means, controlling for Median SAT, Carnegie, Sector, Expenditure, and Size, var (u_{0j})=$\tau_{00.}$ but this is now a conditional or residual variance $\tau_{00|w}$.

R^2 Calculation Formula. The R^2 values were estimated for Level 1 and Level 2 by using Raudenbush and Bryk's (2002) notation:

Level 1: $$R^2_{L1} = \frac{\sigma^2_{\text{unconditional}} - \sigma^2_{\text{conditional}}}{\sigma^2_{\text{unconditional}}} \quad (16)$$

Level 2: $$R^2_{L2} = \frac{\tau^2_{\text{unconditional}} - \tau^2_{\text{conditional}}}{\tau^2_{\text{unconditional}}} \quad (17)$$

Where $\sigma^2_{\text{unconditional}}$ represents the Level 1 unconditional variance; $\sigma^2_{\text{conditional}}$ represents the Level 1 conditional variance; and R^2_{L1} is the proportion of Level 1 variance that is explained by covariate(s). Likewise, $\tau^2_{\text{unconditional}}$ represents the Level 2 unconditional variance; $\tau^2_{\text{conditional}}$ represents the Level 2 conditional variance; and R^2_{L2} is the proportion of Level 2 variance that is explained by covariate(s) (Hedges & Hedberg, 2007).

Research Question 4: MDES Calculation

To answer Question 4, the author calculated the minimum detectable effect size (MDES) using the empirical estimated design parameters for various sample size combinations.

MDES without Covariates. The formula for computing the MDES without covariates is presented below:

$$\mathrm{MDES}_{2\mathrm{LCRT}}=\frac{\mathrm{M}_{\mathrm{J}-2}}{\sqrt{\mathrm{J}}}\sqrt{\rho+\frac{1-\rho}{\mathrm{n}}}\sqrt{\frac{1}{\mathrm{P}(1-\mathrm{P})}} \qquad (18)$$

Where n is the number of individuals per cluster (school); J is the total number of clusters; and M is the group effect multiplier, which corresponds to the value of the t-distribution for a two-tailed test with α = 0.05, power = 0.80, equal variances for groups, and J-2 degrees of freedom. If the degrees of freedom are greater than 20, M is approximately 2.8 (Bloom, 1995). *p* is the ICC and P is the proportion of clusters assigned to treatment which we assume to be 0.50.

MDES with Covariates. The formula for computing the MDES with covariates is presented below:

$$\mathrm{MDES}_{2\mathrm{LCRT}}=\frac{\mathrm{M}_{\mathrm{J}-3}}{\sqrt{\mathrm{J}}}\sqrt{(1-\mathrm{R}^2_{\mathrm{L}2})\rho+\frac{(1-{\mathrm{R}_{\mathrm{L}1}}^{2})(1-\rho)}{\mathrm{n}}}\sqrt{\frac{1}{\mathrm{P}(1-\mathrm{P})}} \qquad (19)$$

All parameters in Equation 19 are defined as they were in Equation 18 with the addition of R^2_{L1}, the proportion of variance explained by Level 1covariate(s); and R^2_{L2}, the proportion of variance explained by Level 2 covariate(s).

Specified Sample Sizes for Calculating MDES. In the K-12 literature, it is common to have approximately 40 schools in a two-level CRT. For illustrative purposes, the author calculated the MDES assuming 20, 40, 60 and 80 total universities/colleges in a study. The author assumed a within university/college sample size of approximately 100 students per school as

CAE required that CLA participating schools recruit at least 100 students to ensure higher internal reliability of the test results (CAE, 2009). All MDES calculations will be conducted using PowerUp! (Dong & Maynard, 2013).

Summary

This study expands on previous work done on K-12 design parameters to students' outcomes in higher education. In this study, the author used a set of two-level HLMs (students nested within universities) to estimate ICCs and R^2 for the total CLA outcome, Performance Task outcome, and Analytical Writing outcome. Then the author demonstrated the importance of these design parameters in calculating the MDES for a study using typical sample sizes. In the following chapter, Chapter IV, the author reported the statistical findings based on the methodology section.

CHAPTER IV

RESULTS

This study sought to empirically estimate ICC and R^2 values to help researchers planning large scale cluster randomized trials to test the efficacy of cognitive skills interventions in higher education. In this chapter, the author begins by reviewing the research questions. Then the author presents the descriptive statistics followed by the empirically estimated ICCs and R^2 values. Specifically, the author reports unconditional ICCs and standard error estimates and the proportion of variance explained by student-level covariates (R^2_{L1}) and school-level covariates (R^2_{L2}) for each outcome and sample combination. Finally, the author reports estimates of the MDES with and without covariates for the three outcomes based on the estimated design parameters and reasonable sample sizes. The section ends with a summary of this chapter.

Review of Research Questions

The research questions that were posed for this study were as follows:

1. To what extent do the following outcomes vary across schools:
   A. Task Performance outcome?
   B. Analytical Writing outcome?
   C. The total CLA outcome?
2. To what extent do student-level covariates (i.e., EAA, student demographic variables) explain the variance in the three outcomes?
3. To what extent do institution-level covariate (i.e., Median SAT, sector, Carnegie classification, etc.) explain the variance in the three outcomes?

4. Given the design parameters estimates in questions 1-3 and effect sizes from the literature, what is the sample size necessary for CRTs which aim to test interventions seeking to improve critical thinking and communication skills at colleges/universities?

Descriptive Statistics of Data

The present empirical analysis uses data from the CLA test which was administered to four-year, not-for-profit colleges and universities in the United States (CAE, 2005). The sample seeks to generalize to four-year, not-for-profit colleges and universities in the United States. Table 8 is an adaption of Klein, Benjamine, Shalverson, and Bolus' Four-Year Institutions in the CLA and Nation by Key Characteristics (2007), which shows a close correspondence between the characteristics of the approximately 1,400 institutions in the IPEDs database and the characteristics of the over 100 schools participating in the CLA (Klein, Benjamine, Shalverson, & Bolus, 2007). As can be seen that only 37 universities in this study since it was drawn from a longitudinal CLA study provided by CAE. However, participating schools in this study appear to be more selective than full set of schools that participate in the CLA. Universities in this study data set tended to have higher Median SAT than the full CLA sample (1,150 vs. 1,079) higher mean four-year graduation (50% vs 38%), first-year retention rates (85% vs 77%), mean number of FTE students (10,000 vs. 6,160), and mean student related expenditure per FTE ($15,001 vs. 11,820).

Table 8
Characteristics of Four-Year Colleges/Universities Samples in This Study

School Level Characteristics	Nation-wide Universities	All CLA Participating Universities	CLA Participating in This Study
Percent of public schools	36%	42%	51%
Percent of HBCU	6%	10%	11%
Mean percentage of Pell Grant receivers	33%	32%	30%
Mean four-year graduation rate	36%	38%	50%
First-year retention rate	75%	77%	85%
Mean six-year graduation rate	52%	55%	51%
Mean median SAT	1,061	1,079	1,150
Mean number of FTE Student	4,500	6,160	10,000
Mean student related expenditure per FTE	$12,230	$11,820	$15,001

Note. HBCU= Historically Black College or University. The table was an adaption of Klein, Benjamine, Shalverson, and Bolus' Four-Year Institutions in the CLA and Nation by Key Characteristics (2007) with a fourth column being added to the original table.

Table 9 displays the demographic data for the sample used for this study. The total number of participants in the sample was 9,827. 3,388 (34.5%) of those were male and 5,968 (60.7%) were female. Table 9 also shows that more than half of participants (5,635) were White students, which accounted for 57.3%, followed by 1,565 (15.9%) Black, 804 (8.2 %) Hispanic, 680 (6.9%) of Asian/Pacific Islander, together with 1.6% (160) American Indian/Alaska Native. In terms of proportion of English as primary Language at home or not, 81.3% (7,992) participants reported "Yes" whereas 13.9% (1,363) counterparts reported "no response".

Table 10 reviews the sample sizes. Recall that sample A represents freshmen who were tested during Phase 1, which occurred in fall 2005 or fall 2006. Sample B is a subset of these students who were also tested in Phase 2, which occurred in spring 2007 or spring 2008. As such, Sample B has two time points of data. Sample C is again a subset of the full sample who were

Table 9
Demographics Statistic Descriptive

Variable	Category	Frequency	Percent
Gender	Female	5,968	60.73
	Male	3,388	34.48
	Missing	471	4.79
	Total	9,827	100.00
Race	Black	1,565	15.93
	American Indian/Alaska	160	1.63
	Asian/Pacific	680	6.92
	Hispanic	804	8.18
	White	5,635	57.34
	Other	512	5.21
	Missing	471	4.79
	Total	9,827	95.21
English as primary language	No	1,363	13.87
	Yes	7,992	81.33
	Total	9,355	95.20
	Missing	472	4.80
	Total	9,827	100.00

Table 10
Analytical Student and School Sample Sizes by Phase in This Study

Sample	Phase 1 (fall 2005 or fall 2006)		Phase 2 (spring 2007 or spring 2008)		Phase 3 (spring 2009 or spring 2010)	
	n	J	n	J	n	J
Sample A	8,061	37	..	..	..	..
Sample B	2,037	27	2,037	27	..	..
Sample C	930	22	930	22	930	22

Note. n=number of students; J=number of schools; "..”=no participants in specific time. Analytical sample sizes are slightly smaller due to missing data which was handled through listwise deletion. Sample A had 8,061 participants in Phase 1.

tested in Phase 2 (occurred in spring 2008 spring 2007 or spring 2008) and Phase 3 (spring 2009 or spring 2010). Hence Sample C has three time points of data.

Tables 11 and 12 display descriptive statistics organized by Level 1 and Level 2 variables. Level 1 variables included EAA, age, the total CLA, Performance Task, and Analytical Writing

Table 11
Level 1 Variable Descriptive by Samples and Phases

		Phase 1		Phase 2		Phase 3	
Sample	Variables	Mean	*SD*	Mean	*SD*	Mean	*SD*
Level 1							
Sample A	Sex	1.60	0.50	..	..	..	..
(n=9,827)	Race	4.10	1.60	..	..	..	..
	Age	18.30	0.90	..	..	..	..
	EPL	0.90	0.40	..	..	..	..
	EAA	1105.70	189.40	..	..	..	..
	CLA Outcome	1100.70	146.80	..	..	..	..
	PT Outcome	1099.50	184.50	..	..	..	..
	AW Outcome	1087.60	163.30	..	..	..	..
Sample B	Sex	1.64	0.48	*	*	..	..
(n=2,422)	Race	4.09	1.53	*	*	..	..
	EPL	0.84	0.37	*	*		
	Age	18.28	0.62	19.43	0.72	..	..
	EAA	1117.01	188.85	1117.01	188.85	..	..
	CLA Outcome	1118.82	147.26	1150.38	160.01	..	..
	PT Outcome	1120.10	184.41	1162.89	207.75	..	..
	AW Outcome	1106.14	165.15	1132.95	160.76	..	..
Sample C	Sex	1.64	0.48	*	*	*	*
(n=1,064)	Race	4.31	1.35	*	*	*	*
	EPL	0.85	0.36	*	*	*	*
	Age	1147.06	185.75	19.44	0.60	21.98	0.51
	EAA	18.29	0.55	1147.06	185.75	1147.06	185.75
	CLA Outcome	1141.55	147.38	1183.36	157.98	1231.06	159.11
	PT Outcome	1144.79	179.10	1197.52	209.53	1229.77	187.95
	AW Outcome	1132.64	167.01	1165.01	158.06	1227.06	184.48

Note. EPL=English as Primary Language; EAA=Entering Academic Ability; PT=Performance Task outcome; AW=Analytical Writing outcome; "*" indicates the same values as in the later phases given that the characteristics do not change over time.

Table 12
Level 2 Variable Descriptive by Samples and Phases

		Phase 1		Phase 2		Phase 3	
Sample	Variables	Mean	*SD*	Mean	*SD*	Mean	*SD*
Level 2							
Sample A	CC	2.14	0.86	..	..	..	..
(J=37)	Sector	1.51	0.51	..	..	..	..
	Median SAT	4.62	1.38	..	..	..	..
	Mean expenditures	3.32	1.03	..	..	..	..
	Size of Enrollment	2.05	0.82	..	..	..	..
Sample B	CC	2.33	0.73	*	*	..	..
(J=27)	Sector	1.44	0.51	*	*	..	..
	Median SAT	4.48	1.28	*	*	..	..
	Mean expenditures	3.19	1.00	*	*	..	..
	Size of Enrollment	2.19	0.74	*	*	..	..
Sample C	CC	2.36	0.73	*	*	*	*
(J=22)	Sector	1.41	0.50	*	*	*	*
	Median SAT	4.41	1.40	*	*	*	*
	Mean expenditures	3.23	1.02	*	*	*	*
	Size of Enrollment	2.27	0.70	*	*	*	*

Note. CC=Carnegie Classification; Mean expenditures=Mean student-related expenditure; "*" indicates the same values as in the later phases given that the characteristics do not change over time.

outcomes arranged by phase. Table 11 also presents descriptive statistics for the institutional characteristics for the colleges/universities in the samples at the different phases.

Results of Empirical Estimates of Design Parameters

The results of empirical estimates of design parameters are divided into three sections that align with the research questions and methods. The author first presents the unconditional ICCs for the two-level HLMs for Samples A, B, and C in Phase 1, B and C in Phase 2, and C in Phase 3 for each outcome (total CLA, critical thinking, and written communication). Next, the author presents the percent of variance in each outcome explained with the different covariate sets for each sample at the different time phases.

Research Question 1: Unconditional Model 1

This section presents the ICCs and standard error (SE) generated from the unconditional model without covariates in total CLA, Performance Task, and Analytical Writing outcomes.

Unconditional Model 1. Table 13 is divided into three panels and four columns each. From left to right, the first column displays samples by each phase. ICCs and SE are shown in column 2 for the total CLA outcome, in column 3 for Performance Task outcome, and in column 4

Table 13
ICCs for Total CLA, Performance Task, and Analytical Writing Outcomes: Model 1

	Total CLA		PT		AW	
Sample	ICC	SE	ICC	SE	ICC	SE
Phase 1						
Sample A	0.311	0.050	0.203	0.038	0.277	0.047
Sample B	0.305	0.058	0.197	0.043	0.271	0.054
Sample C	0.322	0.066	0.194	0.047	0.293	0.063
Phase 2						
Sample A	..	..	..	..	..	..
Sample B	0.330	0.060	0.228	0.048	0.293	0.056
Sample C	0.352	0.069	0.222	0.052	0.320	0.066
Phase 3						
Sample A	..	..	..	..	..	..
Sample B	..	..	..	..	..	..
Sample C	0.353	0.069	0.260	0.058	0.274	0.060
Mean ICCs	**0.329**		**0.217**		**0.288**	

Note. PT=Performance Task outcome; AW=Analytical Writing outcome; ".." indicates that ICCs are not in specific time; ICC= intraclass correlation; SE=standard error; Sample A combined fall 2005 and fall 2006 (Phase 1) freshmen's achievement outcome data; Sample B combined spring 2007 and spring 2008 (Phase 2) rising juniors' achievement outcome data.

for Analytical Writing outcome. The bottom row produces the mean ICCs for each outcome given all samples across all phases.

Total CLA outcome. The ICC for the total CLA outcome ranges from 0.305 to 0.353 for with the mean ICC of 0.329.

Performance Task outcome. The ICC for Performance Task outcome ranges from 0.194 to 0.260 with the mean ICC of 0.217.

Analytical Writing outcome. The ICC for Analytical Writing outcome ranges from 0.271 to 0.320 with the mean of 0.288.

Research Question 2: Conditional Model 2-4

To determine the explanatory power of covariates in improving design efficiency, the author examined the strength of the following sets of covariates in reducing variance for each of the outcomes (Model 2-4): student-level pretests (EAA), student-level demographics, and school-level covariates.

Model 2 with the Pretest (EAA) Model. EAA was included at student level and aggregated up to school level (Mean EAA) using Model 2. Table 14 shows the variance explained by EAA at each level was reported as well as the mean across all samples and outcomes.

Total CLA outcome. The R^2_{L1} including EAA for total CLA outcome ranges from 0.168 to 0.181 with the mean of 0.175. Whereas R^2_{L2} for total CLA outcome ranges from 0.697 to 0.781 with the mean of 0.726.

Performance task outcome. The R^2_{L1} including EAA for Performance Task outcome ranges from 0.104 to 0.151 at the student level with the mean of 0.141. Whereas R^2_{L2} ranges from 0.774 to 0.874 with the mean of 0.838.

Table 14

R^2 Values Including EAA for Total CLA, Performance Task, and Analytical Writing Outcomes: Model 2

	Total CLA		PT		AW	
Sample	R^2_{L1}	R^2_{L2}	R^2_{L1}	R^2_{L2}	R^2_{L1}	R^2_{L2}
Phase 1						
Sample A	0.178	0.781	0.149	0.874	0.088	0.647
Sample B	0.178	0.729	0.146	0.827	0.088	0.588
Sample C	0.168	0.721	0.143	0.883	0.082	0.580
Phase 2						
Sample A	..	..	..	..	..	..
Sample B	0.181	0.727	0.151	0.834	0.091	0.573
Sample C	0.174	0.697	0.151	0.835	0.070	0.517
Phase 3						
Sample A	..	..	..	..	..	..
Sample B	..	..	..	..	..	..
Sample C	0.171	0.704	0.104	0.774	0.122	0.639
Mean R^2	**0.175**	**0.726**	**0.141**	**0.838**	**0.090**	**0.591**

Note. PT=Performance Task outcome; AW=Analytical Writing outcome; ".." indicates that R^2 are not in specific time.

Analytical writing outcome. The R^2_{L1} including EAA ranges from 0.070 to 0.122 with the mean of 0.090. Whereas the R^2_{L2} ranges from 0.517 to 0.647 with the mean of 0.591.

Model 3 with Demographics Model. Table 15 presents the proportion of variance explained by demographic composite at student- and school-level. In general, demographic variables explain less variation than EAA at both levels.

Total CLA outcome. The R^2_{L1} ranges from 0.024 to 0.049 with a mean of 0.039. Whereas the R^2_{L2} ranges from 0.268 to 0.550 with a mean of 0.367 for the total CLA outcome.

Performance task outcome. The R^2_{L1} ranges from 0.019 to 0.038 at the student level with a mean of 0.028. Whereas the R^2_{L2} ranges from 0.301 to 0.492 with a mean of 0.383 for Performance Task outcome.

Table 15
R^2 Values Including Demographics for Total CLA, Performance Task, and Analytical Writing Outcomes: Model 3

	Total CLA		PT		AW	
Sample	R^2_{L1}	R^2_{L2}	R^2_{L1}	R^2_{L2}	R^2_{L1}	R^2_{L2}
Phase 1						
Sample A	0.038	0.365	0.027	0.410	0.026	0.342
Sample B	0.047	0.376	0.038	0.301	0.022	0.463
Sample C	0.049	0.550	0.036	0.492	0.028	0.597
Phase 2						
Sample A	..	..	..	..	..	..
Sample B	0.044	0.366	0.036	0.384	0.030	0.332
Sample C	0.032	0.276	0.019	0.321	0.028	0.221
Phase 3						
Sample A	..	..	..	..	..	..
Sample B	..	..	..	..	..	..
Sample C	0.024	0.268	0.013	0.392	0.016	0.186
Mean R^2	**0.039**	**0.367**	**0.028**	**0.383**	**0.025**	**0.357**

Note. PT=Performance Task outcome; AW=Analytical Writing outcome; ".." indicates that R^2 are not in specific time.

Analytical writing outcome. The R^2_{L1} for analytical writing outcome ranges from 0.016 to 0. 030 with the mean of 0.025. Whereas the R^2_{L2} ranges from 0.186 to 0.597 with the mean 0.357 for Analytical Writing outcome.

Model 4 with the Pretest (EAA) and Demographics Model. Model 4 included both pretest (EAA) and demographic composite as student level covariates to investigate the predictive power of the covariates together. Table 16 displays the results for each outcome. As expected, the combined set of covariates has the greatest explanatory power.

Total CLA Outcome. The R^2_{L1} for the total CLA outcome ranges from 0.182 to 0.194 with a mean of 0.188. Whereas R^2_{L2} ranges from 0.718 to 0.832 with a mean of 0.757.

Table 16
R^2 Values Including EAA and Demographics for Total CLA, Performance Task, and Analytical Writing Outcomes: Model 4

	Total CLA		PT		AW	
Sample	R^2_{L1}	R^2_{L2}	R^2_{L1}	R^2_{L2}	R^2_{L1}	R^2_{L2}
Phase 1						
Sample A	0.189	0.815	0.153	0.871	0.100	0.758
Sample B	0.191	0.735	0.156	0.802	0.093	0.682
Sample C	0.188	0.916	0.155	0.929	0.098	0.914
Phase 2						
Sample A	..	..	..	..	..	..
Sample B	0.193	0.703	0.158	0.818	0.102	0.548
Sample C	0.186	0.665	0.157	0.823	0.085	0.470
Phase 3						
Sample A	..	..	..	..	..	..
Sample B	..	..	..	..	..	..
Sample C	0.179	0.706	0.107	0.833	0.128	0.599
Mean R^2	**0.188**	**0.757**	**0.148**	**0.846**	**0.101**	**0.662**

Note. PT=Performance Task outcome; AW=Analytical Writing outcome; ".." indicates that R^2 are not in specific time.

Performance Task outcome. The R^2_{L1} ranges from 0.107 to 0.158, with the mean of 0.148. Whereas the R^2_{L2} ranges from 0.665 to 0.916 with the mean of 0.757.

Analytical Writing outcome. The R^2_{L1} ranges from 0.093 to 0.128 with the mean of 0.101. Whereas the R^2_{L2} ranges from 0.470 to 0.914 with the mean of 0.662.

Research Question 3: Conditional Model 5-7

Question 3 sought to investigate to what extent the school-level covariates explain the variability in each of the three outcomes (Model 5-7). In some cases, when administrative data are not available or may be expensive, the use of school characteristics as covariates are another option as that information are publicly accessible on IPEDS website. Model 5, 6 and 7 investigated the effect of school-level covariate on reducing both between-school variances in all samples. Note that within school variance is not explained in these cases since school characteristics only have effect on reducing variance at Level 2.

Model 5 with the Median SAT Model. Model 5 included Median SAT as a proxy of school-level pretest to investigate the effect of the school-level covariate on reducing variance in each outcome. Table 17 displays the Level 2 variance (R^2_{L2}) for each outcome.

Table 17
R^2 Values Including Median SAT Covariate for Total CLA, Performance Task, and Analytical Writing Outcomes: Model 5

	Total CLA	PT	AW
Sample	R^2_{L2}	R^2_{L2}	R^2_{L2}
Phase 1			
Sample A	0.742	0.829	0.617
Sample B	0.729	0.827	0.594
Sample C	0.760	0.925	0.614
Phase 2			
Sample A	..	..	..
Sample B	0.713	0.816	0.572
Sample C	0.727	0.854	0.563
Phase 3			
Sample A	..	..	..
Sample B	..	..	..
Sample C	0.834	0.893	0.762
Mean R^2	**0.751**	**0.857**	**0.620**

Note. PT=Performance Task outcome; AW=Analytical Writing outcome; ".." indicates that R^2 are not in specific time.

The total CLA outcome. The R^2_{L2} for median SAT in the total CLA outcome ranges from 0.713 to 0.834 with the mean of 0.751.

Performance Task outcome. The R^2_{L2} for median SAT in the Performance Task outcome ranges from 0.816 to 0.925 with the mean of 0.857.

Analytical Writing outcome. The R^2_{L2} for median SAT in Analytical Writing outcomes ranges from 0.563 to 0.762 with the mean of 0.620.

Model 6 with Institutional Characteristics Model. In Model 6, the author investigated how much school characteristics (e.g., Carnegie classification, sector, size, and mean student-related expenditure) contributed to reducing the school-level variance. Table 18 displays the Level-2 variance (R^2_{L2}) of each outcome.

Table 18
R^2 Values Including Institutional Characteristics for Total CLA, Performance Task, and Analytical Writing Outcomes: Model 6

	Total CLA	PT	AW
Sample	R^2_{L2}	R^2_{L2}	R^2_{L2}
Phase 1			
Sample A	0.514	0.503	0.499
Sample B	0.487	0.430	0.504
Sample C	0.510	0.643	0.407
Phase 2			
Sample A	..	..	..
Sample B	0.391	0.451	0.325
Sample C	0.572	0.724	0.405
Phase 3			
Sample A	..	..	..
Sample B	..	..	..
Sample C	0.679	0.648	0.708
Mean R^2	**0.525**	**0.567**	**0.475**

Note. PT=Performance Task outcome; AW=Analytical Writing outcome; "..”Indicates that data is unavailable for specific period.

The total CLA outcome. The $R^2{}_{L2}$ for institutional characteristics for the total CLA outcome ranges from 0.391 to 0.679 with the mean of 0.525.

***Performance Task outcome*.** The $R^2{}_{L2}$ for institutional characteristics in Performance Task outcome ranges from 0.430 to 0.724 with the mean of 0.567.

***Analytical Writing outcome*.** The R^2_{L2} for institutional characteristics for Analytical Writing outcome ranges from 0.325 to 0.708 with the mean of 0.475.

Model 7 with Median SAT and Institutional Characteristics Model. Model 7 incorporate**s** Median SAT and the institutional characteristics. Table 19 displays the $R^2{}_{L2}$ for each outcome.

Table 19
R^2 Values Including Institutional Characteristics and Median SAT for Total CLA, Performance Task, and Analytical Writing Outcomes: Model 7

	Total CLA	PT	AW
Sample	R^2_{L2}	R^2_{L2}	R^2_{L2}
Phase 1			
Sample A	0.819	0.877	0.727
Sample B	0.796	0.852	0.703
Sample C	0.796	0.989	0.626
Phase 2			
Sample A	..	..	..
Sample B	0.699	0.844	0.527
Sample C	0.755	0.939	0.546
Phase 3			
Sample A	..	..	..
Sample B	..	..	..
Sample C	0.871	0.933	0.835
Mean R2	**0.789**	**0.906**	**0.661**

Note. PT=Performance Task outcome; AW=Analytical Writing outcome; "..”Indicates that data is unavailable for specific period.

Total CLA Outcome. The R^2_{L2} associated with median SAT and institutional characteristics for the CLA outcome ranges from 0.699 to 0.871 with a mean of 0.789.

Performance Task outcome. The R^2_{L2} associated with median SAT and institutional characteristics for the Performance Task outcome ranges from 0.844 to 0.987 with the mean of 0.906.

Analytical Writing outcome. The R^2_{L2} associated with median SAT and institutional characteristics for Analytical Writing outcome ranges from 0.527 to 0.835 with the mean of 0.661.

Research Question 4: MDES with and without Covariates

Tables 20 to 25 present estimates of the MDES obtained from equation 18 and 19 given the estimated values of the ICC, R^2_{L1}, and R^2_{L2} across outcome and grades for samples of 20, 40, 60, and 80 schools with assuming 100 students per school.

MDES without Covariates: Unconditional Model 1. Table 20 shows the MDES without considering covariates based on the ICCs generated from Table 13. As expected, as the school sample size (J) increases, the MDES decreases for each outcome. Stated differently, the more schools in a 2-level CRT, the greater the precision of the study. For example, for total CLA outcome, the mean MDES is 0.738, 0.522, 0.462, and 0.369 for 20, 40, 60, and 80 schools.

MDES with Covariate Model 2. Table 21 shows the MDES results if EAA is included as a covariate. Note that the estimates of ICCs are those from Table 13; R^2_{L1} and R^2_{L2} are from Table 14. The findings suggest that the magnitude of MDES including student-level pretest was reduced to almost half of MDES without covariate. For example, the mean MDES for the total CLA outcome were 0.394 for 20 schools, 0.279 for 40 schools, 0.227 for 60 schools, and 0.197 for 80 schools, respectively.

Table 20

Mean MDES for Total CLA, Performance Task, and Analytical Writing Outcomes: Model 1 (n=100)

	Total CLA(J)				PT(J)				AW(J)			
Sample	20	40	60	80	20	40	60	80		40	60	80
Phase 1												
Sample A	0.719	0.508	0.415	0.359	0.585	0.414	0.338	0.293	0.680	0.480	0.392	0.340
Sample B	0.712	0.503	0.411	0.356	0.577	0.408	0.333	0.289	0.672	0.475	0.388	0.336
Sample C	0.731	0.517	0.422	0.365	0.573	0.405	0.331	0.286	0.698	0.494	0.403	0.349
Phase 2												
Sample A	..	..	..	..	..	..	..	..	..	..	..	..
Sample B	0.740	0.523	0.427	0.370	0.619	0.438	0.357	0.309	0.698	0.494	0.403	0.349
Sample C	0.763	0.539	0.440	0.381	0.611	0.432	0.353	0.305	0.729	0.515	0.421	0.364
Phase 3												
Sample A	..	..	..	..	..	..	..	..	..	..	..	..
Sample B	..	..	..	..	..	..	..	..	..	..	..	..
Sample C	0.764	0.540	0.441	0.382	0.659	0.466	0.381	0.330	0.68	0.48	0.390	0.34
Mean MDES	**0.738**	**0.522**	**0.426**	**0.369**	**0.604**	**0.427**	**0.349**	**0.302**	**0.692**	**0.489**	**0.400**	**0.346**

Note. PT=Performance Task outcome; AW=Analytical Writing outcome; J=school sample size; n=student sample sizes; ".." indicates no values for MDES.

Table 21

MDES Including EAA for Total CLA, Performance Task, and Analytical Writing Outcomes: Model 2 (n=100)

	Total CLA(J)				PT(J)				AW(J)			
Sample	20	40	60	80	20	40	60	80	20	40	60	80
Phase 1												
Sample A	0.346	0.245	0.200	0.173	0.229	0.162	0.132	0.115	0.412	0.291	0.238	0.206
Sample B	0.379	0.268	0.219	0.189	0.258	0.182	0.149	0.129	0.438	0.310	0.253	0.219
Sample C	0.394	0.278	0.227	0.197	0.219	0.155	0.127	0.110	0.459	0.324	0.265	0.229
Phase 2												
Sample A	..	..	..	..	..	..	..	..	..	..	..	..
Sample B	0.394	0.279	0.227	0.197	0.269	0.190	0.155	0.134	0.462	0.327	0.267	0.231
Sample C	0.427	0.302	0.246	0.213	0.265	0.187	0.153	0.133	0.511	0.361	0.295	0.256
Phase 3												
Sample A	..	..	..	..	..	..	..	..	..	..	..	..
Sample B	..	..	..	..	..	..	..	..	..	..	..	..
Sample C	0.422	0.299	0.244	0.211	0.326	0.230	0.188	0.163	0.414	0.292	0.239	0.207
Mean MDES	**0.394**	**0.279**	**0.227**	**0.197**	**0.261**	**0.184**	**0.151**	**0.131**	**0.449**	**0.318**	**0.260**	**0.225**

Note. PT=Performance Task outcome; AW=Analytical Writing outcome; J=school sample size; n=student sample sizes; “..” indicates no values for MDES.

MDES with Covariate Model 3. Now consider another scenario when demographic characteristics are used alone as covariates. Table 22 presents the estimates of MDES for the three outcomes based on estimates of ICCs from Table 13; R^2_{L1} and R^2_{L2} are from Table 15. In this case, the mean MDES for the total CLA outcome were 0.590, 0.417, 0.340, and 0.295 for 20, 40, 60, and 80 schools, respectively.

MDES with Covariate Model 4. Table 23 presents MDES based on estimated ICC in Table 13, R^2_{L1} and R^2_{L2} yielded in Table 16 when student-level pretest EAA and demographic variables were considered. The mean MDES for the total CLA outcome were 0.367, 0.260, 0.212, and 0.184 for 20, 40, 60, and 80 schools, respectively.

MDES with Covariate Model 5. Table 24 presents estimates of MDES when including Median SAT as a proxy for school-level pretest. Note that ICC produced in Table 13 and R^2_{L2} produced in Table 17 was included in calculating MDES. For the total CLA outcome, the mean MDES with school-level pretest were: 0.378 for 20 school, 0.267 for 40 schools, 0.218 for 60 schools, and 0.189 for 80 schools, respectively.

MDES with Covariate Model 6. Table 25 displays the MDES based on the estimate of ICC in Table 13 and R^2_{L2} produced in Table 18, which included institutional characteristics as covariates at Level 2. In general, institutional characteristics have less improvement in precision of study design compared with MDES with Median SAT as covariate included at Level 2. For instance, the mean MDES for the total CLA outcome are: 0.511, 0.361, 0.295, and 0.256 for 20, 40, 60, and 80 schools, respectively.

MDES with Covariate Model 7. The last scenario come into consideration is to include median SAT in conjunction with institutional variables. Table 26 displays the MDES based on

Table 22

MDES Including Demographics for Total CLA, Performance Task, and Analytical Writing Outcomes: Model 3 (n=100)

	Total CLA(J)				PT(J)				AW(J)			
Sample	20	40	60	80	20	40	60	80	20	40	60	80
Phase 1												
Sample A	0.576	0.407	0.332	0.288	0.455	0.322	0.263	0.228	0.555	0.392	0.320	0.277
Sample B	0.566	0.400	0.327	0.283	0.486	0.344	0.281	0.243	0.498	0.352	0.288	0.249
Sample C	0.496	0.351	0.286	0.248	0.416	0.294	0.240	0.208	0.451	0.319	0.260	0.225
Phase 2												
Sample A	..	..	..	..	..	..	..	..	..	..	..	..
Sample B	0.592	0.418	0.342	0.296	0.504	0.356	0.291	0.252	0.574	0.406	0.331	0.287
Sample C	0.651	0.461	0.376	0.326	0.507	0.359	0.293	0.254	0.645	0.456	0.372	0.322
Phase 3												
Sample A	..	..	..	..	..	..	..	..	..	..	..	..
Sample B	..	..	..	..	..	..	..	..	..	..	..	..
Sample C	0.656	0.464	0.379	0.328	0.518	0.367	0.299	0.259	0.611	0.432	0.353	0.306
Mean MDES	**0.590**	**0.417**	**0.340**	**0.295**	**0.481**	**0.340**	**0.278**	**0.241**	**0.556**	**0.393**	**0.321**	**0.278**

Note. PT=Performance Task outcome; AW=Analytical Writing outcome; J=school sample size; n=student sample sizes; “..” indicates no values for MDES.

Table 23

MDES Including EAA and Demographics for Total CLA, Performance Task, and Analytical Writing Outcomes: Model 4

	Total CLA(J)				PT(J)				AW(J)			
Sample	20	40	60	80	20	40	60	80	20	40	60	80
Phase 1												
Sample A	0.320	0.226	0.185	0.160	0.231	0.164	0.134	0.116	0.346	0.244	0.200	0.173
Sample B	0.375	0.265	0.216	0.187	0.273	0.193	0.157	0.136	0.388	0.275	0.224	0.194
Sample C	0.230	0.163	0.133	0.115	0.183	0.129	0.106	0.091	0.226	0.160	0.131	0.113
Phase 2												
Sample A	..	..	..	..	..	..	..	..	..	..	..	..
Sample B	0.410	0.290	0.237	0.205	0.279	0.197	0.161	0.140	0.475	0.336	0.274	0.237
Sample C	0.447	0.316	0.258	0.224	0.273	0.193	0.158	0.136	0.534	0.378	0.309	0.267
Phase 3												
Sample A	..	..	..	..	..	..	..	..	..	..	..	..
Sample B	..	..	..	..	..	..	..	..	..	..	..	..
Sample C	0.421	0.298	0.243	0.210	0.285	0.202	0.165	0.143	0.434	0.307	0.251	0.217
Mean MDES	**0.367**	**0.260**	**0.212**	**0.184**	**0.254**	**0.180**	**0.147**	**0.127**	**0.401**	**0.283**	**0.232**	**0.200**

Note. PT=Performance Task outcome; AW=Analytical Writing outcome; J=school sample size; n=student sample sizes; ".." indicates no values for MDES.

Table 24
MDES Including Median SAT for Total CLA, Performance Task, and Analytical Writing Outcomes: Model 5

	Total CLA(J)				PT(J)				AW(J)			
Sample	20	40	60	80	20	40	60	80	20	40	60	80
Phase 1												
Sample A	0.376	0.266	0.217	0.188	0.263	0.186	0.152	0.132	0.429	0.303	0.248	0.215
Sample B	0.382	0.270	0.220	0.191	0.262	0.185	0.151	0.131	0.437	0.309	0.252	0.218
Sample C	0.370	0.261	0.213	0.185	0.192	0.136	0.111	0.096	0.442	0.312	0.255	0.221
Phase 2												
Sample A	..	..	..	..	..	..	..	..	..	..	..	..
Sample B	0.406	0.287	0.234	0.203	0.284	0.201	0.164	0.142	0.464	0.328	0.268	0.232
Sample C	0.408	0.289	0.236	0.204	0.256	0.181	0.148	0.128	0.488	0.345	0.282	0.244
Phase 3												
Sample A	..	..	..	..	..	..	..	..	..	..	..	..
Sample B	..	..	..	..	..	..	..	..	..	..	..	..
Sample C	0.325	0.230	0.188	0.163	0.239	0.169	0.138	0.120	0.343	0.243	0.198	0.172
Mean MDES	**0.378**	**0.267**	**0.218**	**0.189**	**0.249**	**0.176**	**0.144**	**0.125**	**0.434**	**0.307**	**0.251**	**0.217**

Note. PT=Performance Task outcome; AW=Analytical Writing outcome; J=school sample size; n=student sample sizes; ".." indicates no values for MDES.

Table 25

MDES Including Institutional Characteristics for Total CLA, Performance Task, and Analytical Writing Outcomes: Model 6

	Total CLA(J)				PT(J)				AW(J)			
Sample	20	40	60	80	20	40	60	80	20	40	60	80
Phase 1												
Sample A	0.507	0.358	0.293	0.253	0.421	0.297	0.243	0.210	0.487	0.344	0.281	0.244
Sample B	0.515	0.364	0.297	0.258	0.442	0.313	0.255	0.221	0.480	0.339	0.277	0.240
Sample C	0.517	0.366	0.299	0.259	0.354	0.251	0.205	0.177	0.542	0.383	0.313	0.271
Phase 2												
Sample A	..	..	..	..	..	..	..	..	..	..	..	..
Sample B	0.581	0.411	0.335	0.290	0.465	0.329	0.268	0.232	0.577	0.408	0.333	0.288
Sample C	0.505	0.357	0.292	0.253	0.335	0.237	0.193	0.167	0.557	0.394	0.321	0.278
Phase 3												
Sample A	..	..	..	..	..	..	..	..	..	..	..	..
Sample B	..	..	..	..	..	..	..	..	..	..	..	..
Sample C	0.441	0.312	0.255	0.221	0.401	0.283	0.231	0.200	0.377	0.266	0.217	0.188
Mean MDES	**0.511**	**0.361**	**0.295**	**0.256**	**0.403**	**0.285**	**0.233**	**0.201**	**0.503**	**0.356**	**0.290**	**0.252**

Note. PT=Performance Task outcome; AW=Analytical Writing outcome; J=school sample size; n=student sample sizes; “..” indicates no values for MDES.

Table 26

MDES Including Institutional Characteristics and Median SAT for Total CLA, Performance Task, and Analytical Writing Outcomes: Model 7

	Total CLA(J)				PT(J)				AW(J)			
Sample	20	40	60	80	20	40	60	80	20	40	60	80
Phase 1												
Sample A	0.320	0.227	0.185	0.160	0.231	0.164	0.134	0.116	0.367	0.259	0.212	0.183
Sample B	0.335	0.273	0.194	0.168	0.246	0.174	0.142	0.123	0.378	0.267	0.218	0.189
Sample C	0.363	0.257	0.210	0.182	0.129	0.091	0.074	0.064	0.435	0.308	0.251	0.218
Phase 2												
Sample A	..	..	..	..	..	..	..	..	..	..	..	..
Sample B	0.415	0.293	0.240	0.208	0.265	0.188	0.153	0.133	0.486	0.344	0.281	0.243
Sample C	0.388	0.274	0.224	0.194	0.186	0.132	0.107	0.093	0.497	0.351	0.287	0.249
Phase 3												
Sample A	..	..	..	..	..	..	..	..	..	..	..	..
Sample B	..	..	..	..	..	..	..	..	..	..	..	..
Sample C	0.291	0.206	0.168	0.145	0.201	0.142	0.116	0.100	0.292	0.206	0.169	0.146
Mean MDES	**0.352**	**0.255**	**0.204**	**0.176**	**0.210**	**0.149**	**0.121**	**0.105**	**0.409**	**0.289**	**0.236**	**0.205**

Note. PT=Performance Task outcome; AW=Analytical Writing outcome; J=school sample size; n=student sample sizes; "..” indicates no values for MDES.

the estimate of ICC in Table 13 and R2L2 produced in Table 19. For example, MDES for the total CLA outcome are: 0.352, 0.255, 0.204, and 0.176 for 20, 40, 60, and 80 schools, respectively.

Summary

This chapter presented the descriptive statistics for all the sample sizes. To answer Question 1, the author presents the empirically estimated values of ICCs for the total CLA, Performance Task, and Analytical Writing outcomes. To answer Question 2-3, the author presented the R^2 for the different covariate sets for each outcome. Finally, to answer Question 4, the author displayed estimates the MDES for each outcome under different sample size scenarios using the estimated design parameters from questions 1 and 2. In Chapter V, a summary of findings was presented along with the implications for the field.

CHAPTER V

DISCUSSION

Remarkable progress has been made in building a repository of empirical estimates of ICCs and R^2 for student achievement outcomes in K-12 settings. However, there is still limited information available on these design parameters for studies focused on higher education. Without these design parameters, it is challenging to conduct accurate *a priori* power analyses. To that end, the main goal of the current study was two-fold: (1) to provide empirical estimates of ICCs and R^2 to improve the planning and power analyses for researchers planning intervention studies focused on improving cognitive skills, (2) to demonstrate the application of these design parameters by assessing the MDES of a CRT design under various sample size assumptions. Data from longitudinal CLA tests between 2005 to 2010 were used to calculate the design parameters. The results from this study can directly inform researchers planning trials to identify the effect of cognitive skill interventions in higher education.

Summary of Major Findings

In this section, the author discusses the findings of design parameters and MDES from this study and contrast the results with existing work.

Major Findings of ICCs

The first ICC major finding is that the trend suggests that the ICC is largest for total CLA, then Performance Task, and then Analytical Writing (see Table 27). Specifically, the ICCs for the total CLA outcomes ranged between 0.305 and 0.353 with the mean of 0.329. This suggests that approximately 30.5% to 35.3% of the variance in the total CLA outcome is between schools.

Table 27

ICC Range and Mean ICC for Total CLA, Performance Task Outcome, and Analytical Writing Outcome

Outcomes	ICC Range	Mean ICC
Total CLA	0.305-0.353	0.329
PT	0.194-0.228	0.271
AW	0.271-0.320	0.288

Note. Performance Tasks Outcome=PT; Analytic Writing Task Outcome=AWT.

The next largest ICCs were found in the Analytical Writing outcome, which ranged from 0.271 to 0.320 with the mean of 0.288. This implies that between 27.1% and 32.0% of variance in the Analytical Writing outcome is between schools. The smallest ICCs were found in Performance Task outcome which ranged from 0.194 to 0.228 with the mean of 0.217. This implies that between 19.4 % and 22.8% of the variance in the Performance Task outcome is between schools. These findings related to the ICCs were consistent with the current literature on the variability between institutions in higher education students' cognitive skills (Hu & Kuh, 2003; Kim, 2001; Kinzie, Thomas, Palmer, Umbach, & Kuh, 2007; McCormick, Kuh, Pike, & Chen, 2009; Liu, 2009; Steedle, 2012).

The second major finding related to ICCs is that within an outcome and sample, the ICCs are quite consistent across phase 1 and phase 2, though in some cases there appears to be some potential differences at phase 3. Note that Sample A was not applicable for this discussion as it was only measured during Phase 1. The author begins with Sample B. Figure 5 shows the ICCs for Sample B for each outcome across the two phases. Note that ICCs for Sample C in Phase 1 is denoted by solid blue bars and ICCs in Phase 2 is denoted by blank bars. The ICC for the total

CLA in Phase 1 (e.g., 0.305) is similar to the ICCs in Phase 2 (e.g., 0.330). A similar pattern was found in across phases for the Performance Task outcome and the Analytical Writing outcome.

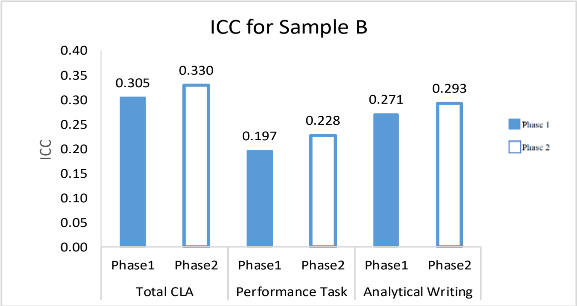


Figure 5. ICCs for Sample B by Phase

Next, the author considers Sample C. In Figure 6, the ICCs for Sample C in Phase 1 is denoted by solid blue bars, the ICCs in Phase 2 is denoted by blank bars, and the ICCs in Phase 3 is denoted by stripe bars. As evident in Figure 6, the ICC trend for the total CLA in Phase 1 is pretty stable across all three phases. That is, the ICC is 0.322 in Phase 1, 0.352 in Phase 2, and 0.353 in Phase 3. However, the ICC trend for the Performance Task outcome trended upwards. That is, the ICC is 0.194 in Phase 1, 0.226 in Phase 2, and then to 0.260 in Phase 3. Further, the ICC trend in Analytical Writing outcome was somewhat inconsistent. That is, the ICC is 0.293 in

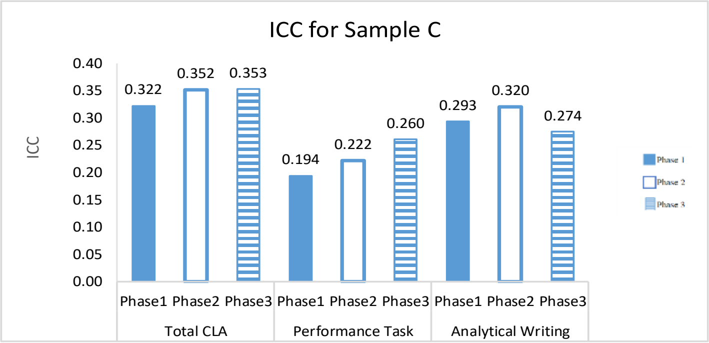


Figure 6. ICCs for Sample C by Phase

Phase 1, then increase slightly to 0.320, and drops to 0.274. Some of these differences across phases may be a result of changes in the makeup of students over four years within a university. However, further qualitative analyses would be helpful to try to uncover why some of these differences may exist.

Major Findings of R^2

In this section, the author summarizes the findings of the estimated values of R^2_{L1} (the proportion of the student-level variance that is predicted by a covariate) and R^2_{L2} (the proportion of the school-level variance that is predicted by a covariate) from six different covariate models.

Student-level covariates. Table 28 compares the mean R^2_{L1} and mean R^2_{L2} from the models with student level covariates including Model 2 (EAA), Model 3 (Demographics only), and Model 4 (EAA and demographics). Overall, the EAA and demographic covariates together have most explanatory power. For this set of covariates, the mean R^2_{L1} was 0.188 and mean R^2_{L2} was 0.757 in total CLA outcome; the mean R^2_{L1} was 0.148 and mean R^2_{L2} was 0.846 in Performance Task outcome; and the mean R^2_{L1} was 0.101 and mean R^2_{L2} was 0.662 in Analytical Writing outcome. However, it is important to note that the explanatory power associated with EAA only was very similar. The mean R^2_{L1} was 0.175 and mean R^2_{L2} was 0.726 in total CLA outcome; the mean R^2_{L1} was 0.141 and mean R^2_{L2} was 0.838 in Performance Task outcome; and the mean R^2_{L1} was 0.090 and mean R^2_{L2} was 0.591 in Analytical Writing outcome. This suggests that the inclusion of EAA is the key driver in reducing variation, not the inclusion of the demographics. This is further confirmed by the fact that the explanatory power of the demographics alone was much smaller. Specifically, the mean R^2_{L1} was 0.039 and mean R^2_{L2} was 0.367 in total CLA outcome; the mean R^2_{L1} was 0.028 and mean R^2_{L2} was 0.383 in Performance Task outcome; and the mean R^2_{L1} was 0.025 and mean R^2_{L2} was 0.357 in Analytical Writing outcome.

Collectively, R2 range is closely to published variance Steedle (2012); 0.03-0.06 (R^2_{L1}). 0.87-0.95(R^2_{L2}).

Table 28

Mean R^2 Based on Student-Level Covariates

Models	Total CLA		PT		AW	
	Mean R^2_{L1}	Mean R^2_{L2}	Mean R^2_{L1}	Mean R^2_{L2}	Mean R^2_{L1}	Mean R^2_{L2}
Model 2(EAA)	0.175	0.726	0.141	0.838	0.090	0.591
Model 3(Demographic)	0.039	0.367	0.028	0.383	0.025	0.357
Model 4(EAA and demographic)	0.188	0.757	0.148	0.846	0.101	0.662

Note. PT=Performance Task outcome; AW=Analytical Writing outcome.

This finding is very similar to the K-12 literature which reveals similar trends which suggest that the pretest explains much more variance in the outcome than demographic characteristics.

School-level covariates. Table 29 summarizes the findings from the models that only include school level variables. Note that the values of $R^2\ _{L1}$ are zero given the fact that these are school level covariates and hence they cannot reduce variation at the student level (Bloom et al., 2005).

The most effective covariates for explaining variation in the outcome were from Model 7 (median SAT and institutional covariate model). When considering median SAT and institutional characteristics, the mean R^2_{L2} was 0.789 for the total CLA outcome, 0.906 for the Performance Task, and 0.661 for the Analytical Writing, respectively. However, similar to the findings for the student level covariates, Model 5 (median SAT) which is a proxy for a pretest, yielded similar R^2_{L2} values. The mean R^2_{L2} was 0.751 for the total CLA outcome, 0.857 for the Performance

Task, and 0.620 for the Analytical Writing. Model 3 (institutional covariates only) explained the least amount of variance with a mean R^2_{L2} was 0.525 for the total CLA outcome, 0.567 for the Performance Task, and 0.475 for the Analytical Writing. These findings suggest that the inclusion of median SAT is more important in explaining variance in these outcomes than institutional characteristics, which is similar to the findings from the student level covariates in K-12 literature.

Table 29
Mean R^2 Based on School-Level Covariates

Models	Total CLA	PT	AW
	Mean R^2_{L2}	Mean R^2_{L2}	Mean R^2_{L2}
Model 5(Median SAT)	0.751	0.857	0.620
Model 6(Institutional)	0.525	0.567	0.475
Model 7 (Median SAT and Institutional)	0.789	0.906	0.661

Note. PT=Performance Task outcome; AW=Analytical Writing outcome.

In conclusion, covariates at both the student and/or school levels have advantages in reducing the within and between school variance of students' outcomes. The findings from this study are consistent with past research in K-12 that using a pre-test at either the student or school level can explain a larger proportion of the outcome variance and hence dramatically increase the precision of a study (Bloom, et al., 1999; Bloom et al., 2005; Hedges & Hedberg, 2007). A school level covariate, such as median SAT is often readily available from the IPEDS website which will make it much easier and cost-effective to obtain than a student level pretest. Given that the explanatory power of median SAT is similar at the school level to the explanatory power of a student level pretest and the fact that reducing the variance at the school level is critical in increasing the precision of a study, the median SAT is likely a good choice to include in a model. Student level demographics and school-level institutional characteristics can also help explain variance in the outcomes, though they do not tend to be as powerful as the pretests.

Major Findings of MDES

This section discusses the findings of MDES based on corresponding design parameters reported above for each outcome. Table 30 compares the summary of mean MDESs for the seven models arranged in a low-to-high order. Recall that the smaller the MDES, the more precise of an estimate of the treatment effect of a CRT study is. As can be seen in Table 29, the precision of a CRT study is improved substantially by including median SAT and institutional variables, EAA and demographic composite or median SAT alone, which can reduce variance in all the three outcomes.

To summarize, the MDES calculations in this study revealed the following. In the case that no covariates are available and assuming 100 individuals per university, 80 total schools were necessary to yield a MDES in the range of 0.302 to 0.369 across the three outcomes. The inclusion of covariates, in particular either the student or school level pretest, greatly reduced the variance in the outcomes and thus increase the power of the study to detect a treatment effect. For example, the MDES ranged from 0.105 to 0.295 across the three outcomes for a total of 80 clusters. This represents great gains in precision from the case without covariates which further strengthens arguments for the importance of including covariates in planning CRTs to test the impact of higher education interventions.

Table 30

Mean MDES (low-high order) By Models

	Total CLA				PT				AW			
	J				J				J			
Model	20	40	60	80	20	40	60	80	20	40	60	80
Model 7 (median SAT and institutional characteristics)	0.352	0.255	0.204	0.176	0.210	0.149	0.121	0.105	0.409	0.289	0.236	0.205
Model 4 (EAA and Demographics)	0.367	0.260	0.212	0.184	0.254	0.180	0.147	0.127	0.401	0.283	0.232	0.200
Model 5 (median SAT)	0.378	0.267	0.218	0.189	0.249	0.176	0.144	0.125	0.434	0.307	0.251	0.217
Model 2 (EAA)	0.394	0.279	0.227	0.197	0.261	0.184	0.151	0.131	0.449	0.318	0.260	0.225
Model 6 (Institutional characteristics)	0.511	0.361	0.295	0.256	0.403	0.285	0.233	0.201	0.503	0.356	0.290	0.252
Model 3(Demographics)	0.590	0.417	0.340	0.295	0.481	0.340	0.278	0.241	0.556	0.393	0.321	0.278
Model 1(ANOVA)	0.738	0.522	0.426	0.369	0.604	0.427	0.349	0.302	0.692	0.489	0.400	0.346

Note . PT=Performance Task outcome; AW=Analytical Writing outcome; J=school sample size; “..” indicates no values for MDES.

Implications of the Study Findings

Recently, higher education has placed a premium on rigorous research to test the impact of educational interventions. This may be due in part to concerns about a deficiency of high-quality evidence of the effectiveness of pedagogical approaches and curriculum redesign to improve undergraduate students' cognitive skills (Dehar-Horenstein & Liu, 2011; Tiruneh, Verburgh & Elen, 2013). In response to the call, the author anticipates that there will be an increase in the design of CRTs to test interventions designed to improve undergraduates' cognitive skills and it is critical that these CRTs are designed with adequate power to detect a meaningful treatment effect.

To design such an impact research of cognitive skill interventions, one must use estimates of design parameters to calculate the MDES. The goal of this study is to provide relevant estimates of these design parameters. For example, suppose a team of researchers are planning a CRT to test the impact of an intervention cognitive skill using the total CLA outcome, the researchers need an estimate of the ICC and relevant R^2 of the total CLA to conduct the power analysis. Further, imagine that they plan to use median SAT as a covariate to increase the precision of the estimate. If they have 100 students per university and either 20, 40, 60, or 80 total universities in their study, they could go directly to the Table 29 in this dissertation to determine the MDES: 0.378, 0.267, 0.218, and 0.189, respectively. However, it is often the case that they will have a different number of total universities. In that case, they could use the design parameters in Table 13-19 in this dissertation to estimate the ICC and R^2. Then they could go to PowerUp! or any other statistical power software to compute the MDES. It is also important to note that the computed MDES must then be examined to determine if it is reasonable. In K-12, the literature suggests that it is often reasonable to design a CRT to test an intervention aimed at improving academic outcomes to detect an effect size of 0.20. Whereas in higher education literature, it is plausible to design a CRT to test an intervention aimed at improving cognitive skills to detect an effect size depending on measures and treatment intensity (Arum, Roksa, & Cho, 2011; Huber et al., 2016; Niu, Behar-Horenstein, & Garven, 2013; Ortiz, 2007; Pascarella &Terenzini, 2005; Pascarella, Blaich, Martin, & Hanson, 2011).

Limitations and Delimitations

To my knowledge, this is the first compilation of empirically estimated values of ICCs and R^2 for cognitive skill outcomes for students in higher education. Although the findings from the study are useful for specific contexts, it is important to consider the limitations of the findings.

First, ICCs are sensitive to specific samples hence researchers should think carefully about whether the samples in the studies they are planning are similar (Kelcey & Phelps, 2013). In this study, the sample are four-year, not-for-profit colleges and universities. Therefore, it is upon researchers to consider carefully to what extent the samples involved in the CRT they are planning are similar to those in this study.

The second limitation is that the outcomes in this study are limited to the cognitive skills domain. As we know from the K-12 literature, the empirical estimates of design parameters may not transfer to other outcome domains for several reasons. First, ICCs for Performance Task, Analytical Writing, and the total CLA outcomes vary by samples, domain, and phases. For example, the reported Performance Task ICC is 0.197 (see Table 13) for Sample B in Phase 1, and 0.228 for Sample B in Phase 3. The slight difference could have an influence on sample sizes.

A third limitation is that due to high attrition in the longitudinal study dataset, many student records in Phase 2 and 3 were not available for analysis, which potentially introduces a certain amount of bias in the design parameters results. Sensitivity analyses were performed to assess the robustness of the ICCs results based on primary analyses of data. The results suggest that the missing data did not bias the results, but it should be noted that there were large amounts of missing data in Phase 2 and 3.

The last limitation lies in the choice of covariates. The choice of covariates was limited to those that were included in the CLA dataset or the IPEDS dataset hence other covariates that may have explained variation in the outcome, such as HSGPA, one-year lagged CLA tests as pretests, interaction with faculty, and motivations, just name a few.

The study also has delimitation bounds. The focus of this study was on the use of the estimated design parameters to plan two-level CRTs but results also can apply to two-level quasi-experimental designs (Spybrook, Westin, &Taylor, 2013). Although for the MDES calculations also focused on balanced designs, or an equal number of clusters per conditions, the calculations could also be extended to unbalanced designs. The design parameters could also be used to help plan another type of CRT design, a two-level blocked CRT design (Level 2 is the unit of random assignment) (Konstantopolous, 2012; Dong, et al., 2016).

Recommendations for Future Research

This study serves as the beginning of a compendium of design parameters for planning impact studies focusing on cognitive skill intervention in higher education. But it is important to recognize that several steps are necessary to help advance the progress of experimental studies in this area for future research.

First, as suggested by Niehaus et al. (2013), researchers in higher education should also report ICCs and percentage of variance explained from covariate sets at each level as part of routine practice. Specifically, researchers are advised to report from three dimensions as the CONSORT guideline developed by Campell et al. (2004): (1) description of the datasets and outcomes; (2) information on the calculation of ICCs; and (3) information on the precision of ICCs. This will help the design parameter database in higher education to continue to expand. Beyond that, publishing effect sizes to help researchers assess whether an MDES is reasonable or not is also important.

Second, researchers are encouraged to continuously add to the design parameters database in higher education by expanding to other outcomes and populations. For example, other outcome measures may include the Watson–Glaser Critical Thinking Appraisal (WGCTA, Watson & Glaser, 1980) and Cornell Critical Thinking Tests (CCTT, Ennis & Millman, 1985), and California Critical Thinking Skills Test (CCTST, Facione, 1990a). The WGCTA, CCTT, and CCTST also target other types of populations including psychology and nursing students. Expanding to other outcomes and populations would help expand the use of the database as it would make it more relevant for other types of interventions that are often tested in higher education including various forms of discussion (Daud & Husin, 2004; Elliot et al., 2001; Garside, 1996; Stark, 2012; Szabo & Schwartz, 2011; Yang & Chou, 2008), concept maps and argument diagrams (Bonk & Smith, 1998; Lee, et al., 2011; Wheeler & Collins, 2003;Van Gelder, 2005), and Problem Based Learning (Bonk & Smith, 1998; Norman & Schmidt, 2000; Schmidt, 1983), among others.

Third, future studies can extend ICCs and R^2 for studies with more than two levels of nesting. For example, previous studies have found that fields of study in college differ in degree to which they contribute to growth in reasoning and communication skills as measured by the CLA test (Arum & Roksa, 2008; Klein et al, 2008; Shavelson, 2009; Steedle & Bradley, 2012). Thus, the design parameters in cognitive skill domain can extend to a three-level models (students nested in fields of study, fields of study nested in institutions). In addition, undergraduate students may be nested within sub-clusters within universities. These sub-clusters may be first-year seminars and experience, writing intensive courses, learning communities, etc., which are designed to improve student academic outcomes through "high impact practices or programs" (Austin, 1993; Kuh, 2008). Hence, it also extends this work to estimate design parameters for three level studies with students nested within seminars or courses that are nested within universities would be useful.

As the increased interests in documenting design parameters for planning CRTs expanded to international level dataset such as Asian countries (Zopluoglu, 2012), Sub-Saharan Africa (Kelcey, Shen, & Spybrook, 2016), and Program for International Student Assessment (PISA) dataset covering as extensive as 81 countries (Brunner, Keller, Wenger, Fischbach & Lüdtke , 2017), further studies may consider generating design parameters from the datasets of multi-national Assessment of Higher Education Learning Outcomes (AHELO) Feasibility Study undertaken by the Organization for Economic Co-operation and Development (OECD). More efforts are encouraged to continuously add to the CLA design parameters database by expanding to other populations and settings to benefit researchers in other countries when coming to conduct impact study relevant to cognitive skills interventions.

Summary

This study was motivated by the call to use RCTs to conduct rigorous evaluations of interventions and programs in higher education. The study empirically estimates ICCs and the percent of variance explained by student and school-level covariates on the total CLA, Performance Task, and Analytical Writing outcomes using the CLA data. Researchers planning CRTs to test the efficacy of interventions aimed at increasing these types of outcomes can use the empirical estimates provided in this study to conduct *a priori* power analyses. This study represents a beginning of a collection of design parameters relevant to higher education and extending this work to other outcome domains relevant to higher education would be useful.

REFERENCES

Arum, R., & Rosa, J. (2011). *Academically adrift: Limited learning on college campuses*. School of Chicago Press.

Astin, A. W. (1984). Student involvement: A developmental theory for higher education. *Journal of College Student Personnel*, *25*(2), 297-308.

Astin, A. W., & Denson, N. (2009). Multi-campus studies of college impact: Which statistical method is appropriate? *Research in Higher Education*, *50*(4), 354-367.

Atkinson, R. (2001). Standardized tests and access to American universities. *The 2001 Robert H. Atwell Distinguished Lecture.American Council on Education, Washington, D.C.*

Atkinson, R. C., & Geiser, S. (2009). Reflections on a century of college admissions tests. *Educational Researcher*, *38*(9), 665-676.

Barnett, E. A., Bork, R. H., Mayer, A. K., Pretlow, J., Wathington, H. D., & Weiss, M. J. (2012). *Bridging the Gap: An Impact Study of Eight Developmental Summer Bridge Programs in Texas*. National Center for Postsecondary Research.

Behar-Horenstein, L. S., & Niu, L. (2011). Teaching critical thinking skills in higher education: A review of the literature. *Journal of College Teaching and Learning, 8*(2), 25.

Benjamin, R. (2014). Two questions about critical-thinking tests in higher education. *Change: The Magazine of Higher Learning*, *46*(2), 24-31.

Berger, J. B., & Milem, J. F. (2000). Exploring the impact of historically Black colleges in promoting the development of undergraduates' self-concept. *Journal of College Student Development*, *41*(4), 1.

Bettinger, E. P., & Baker, R. B. (2014). The effects of student coaching: An evaluation of a randomized experiment in student advising. *Educational Evaluation and Policy Analysis, 36*(1), 3-19.

Black, A. C., Harel, O., & McCoach, D. B. (2011). Missing data techniques for multilevel data: Implications of model misspecification. *Journal of Applied Statistics*, *38*(9), 1845-1865.

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, S. W.-Y. (1999). Using cluster random assignment to measure program impacts. *Evaluation Review, 23*(4), 445-469.

Bok, D. (2009). *Our underachieving colleges: A candid look at how much students learn and why they should be learning more*. Princeton University Press.

Bonk, C. J., & Smith, G. S. (1998). Alternative instructional strategies for creative and critical thinking in the accounting curriculum. *Journal of Accounting Education*, *16*(2), 261-293.

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

Borman, G. D. (2002). Experiments for educational evaluation and improvement. *Peabody Journal of Education*, *77*(4), 7-27.

Borman, G. D., Hewes, G., Rachuba, L. T., & Brown, S. (2002). *Comprehensive school reform and student achievement: A meta-analysis* (Crespar Report No. 59). Baltimore: Johns Hopkins University, Center for Research on the Education of Students Placed at Risk.

Borman, G. D., Slavin, R. E., Cheung, A. C., Chamberlain, A. M., Madden, N. A., & Chambers, B. (2007). Final reading outcomes of the national randomized field trial of Success for All. *American Educational Research Journal*, *44*(3), 701-731.

Boruch, R. F. (1997). *Randomized experiments for planning and evaluation: A practical guide* (Vol. 44). Sage.

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.

Bradley, M., & Steedle, J. T. (2012). *Majors matter: Differential performance on a test of general college outcomes*.

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.

Broton, K. M., Goldrick-Rab, S., & Benson, J. (2016). Working for college: The causal impacts of financial grants on undergraduate employment. *Educational Evaluation and Policy Analysis*, *38*(3), 477-494.

Brunner, M., Keller, U., Wenger, M., Fischbach, A., & Lüdtke, O. (2018). Between-School Variation in Students' Achievement, Motivation, Affect, and Learning Strategies: Results from 81 Countries for Planning Group-Randomized Trials in Education. *Journal of Research on Educational Effectiveness*, *11*(3), 452-478.

Callan, P. M., and Finney, J. E. (2002). Assessing educational capital: An imperative for policy. *Change* (*34*): 25–31.

Carini, R. M., Kuh, G. D., & Klein, S. P. (2006). Student engagement and student learning: Testing the linkages. *Research in Higher Education*, *47*(1), 1-32.

Chickering, A. W., & Hannah, W. (1969). The Process of Withdrawal. *Liberal Educ*.

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences*. New York: Academic Press.

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*, 176–198.

Cook, T. D., Campbell, D. T., & Day, A. (2002). *Quasi-experimentation: Design & analysis issues for field settings* (Vol. 351). Boston: Houghton Mifflin.

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

Cook, T. D., Murphy, R. F., & Hunt, H. D. (2000). Comer's School Development Program in Chicago: A theory-based evaluation. *American Educational Research Journal*, *37*(2), 535-597.

Cook, T. D., & Payne, M. R. (2002). Objecting to the objections to using random assignment in educational research. *Evidence matters: Randomized trials in education research*, 150-178.

Cox, B. E., McIntosh, K., Reason, R. D., & Terenzini, P. T. (2014). Working with missing data in higher education research: A primer and real-world example. *The Review of Higher Education*, *37*(3), 377-402.

Cunha, J. M., & Miller, T. (2014). Measuring value-added in higher education: Possibilities and limitations in the use of administrative data. *Economics of Education Review*, *42*, 64-77.

Daud, N. M., & Husin, Z. (2004). Developing critical thinking skills in computer-aided extended reading classes. *British Journal of Educational Technology*, *35*(4), 477-487.

Dedrick, R. F., Ferron, J. M., Hess, M. R., Hogarty, K. Y., Kromrey, J. D., Lang, T. R., & Lee, R. S. (2009). Multilevel modeling: A review of methodological issues and applications. *Review of Educational Research*, *79*(1), 69-102.

Doherty, K. M. (2000). *Early implementation of the Comprehensive School Reform Demonstration (CSRD) program*. Washington, DC: U.S. Department of Education.

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

Donner, A., Brown, K. S., & Brasher, P. (1990). A methodological review of non-therapeutic intervention trials employing cluster randomization, 1979–1989. *International Journal of Epidemiology, 19*(4), 795-800.

Donner, A., & Klar, N. (2000). *Design and analysis of cluster randomization trials in health research* (pp. 6-10).

Earle, J., Maynard, R., Neild, R. C., Easton, J. Q., Ferrini-Mundy, J., Albro, E., & Winter, S. (2013). *Common guidelines for education research and development*. Washington, DC: IES, DOE, and NSF.

Elliott, B., Oty, K., McArthur, J., & Clark, J. (2001). The effect of an interdisciplinary algebra/science course on students' problem solving skills, critical thinking skills and attitudes towards mathematics. *International Journal of Mathematical Education in Science and Technology, 32*(6), 811-816.

Ennis, R. H. (2018). Critical thinking across the curriculum: A vision. *Topoi, 37*(1), 165-184.

Ennis, R. J., & Millman, J. (1985). *Cornell tests of critical thinking*. Pacific Grove, CA: Midwest.

Evans, W. N., Kearney, M. S., Perry, B. C., & Sullivan, J. X. (2017). *Increasing Community College Completion Rates among Low-Income Students: Evidence from a Randomized Controlled Trial Evaluation of a Case Management Intervention* (No. w24150). National Bureau of Economic Research.

Experimental Sites and Initiatives (March, 2018). https://experimentalsites.ed.gov/exp/index.html

Facione, P. A. (1990). The California critical thinking skills test (CCTST): Form A (1990) and form B (1992). Millbrae.

Finn, J. D., & Achilles, C. M. (1990). Answers and questions about class size: A statewide experiment. *American Educational Research Journal*, *27*, 557–577.

Flay, B. R., Biglan, A., Boruch, R. F., Castro, F. G., Gottfredson, D., Kellam, S., & Ji, P. (2005). Standards of evidence: Criteria for efficacy, effectiveness and dissemination. *Prevention Science*, *6*(3), 151-175.

Flay, B. R., & Collins, L. M. (2005). Historical review of school-based for evaluating problem behavior prevention programs. *The Annals of the American Academy of Political and Social Science*, *599*(1), 115-146.

Garside, C. (1996). *Look who's talking: A comparison of lecture and group discussion teaching strategies in developing critical thinking skills*.

Gates, S. M., Augustine, C. H., Benjamin, R., Bikson, T. K., & Derghazarian, E. (2001). *Ensuring the quality and productivity of education and professional development activities: A review of approaches and lessons for DoD* (No. RAND-MR-1257-OSD). RAND CORP SANTA MONICA CA.

Gelder, T. V. (2005). Teaching critical thinking: Some lessons from cognitive science. *College Teaching*, *53*(1), 41-48.

Goldberg, B., & Finkelstein, M. (2002). Effects of a first-semester learning community on nontraditional technical students. *Innovative Higher Education*, *26*(4), 235-249.

Gordon, T. W., Young, J. C., & Kalianov, C. J. (2001). Connecting the freshman year experience through learning communities: Practical implications for academic and student affairs units. *College Student Affairs Journal*, *20*(2), 37.

Graham, J. W. (2012). *Missing data: Analysis and design*. Springer Science & Business Media.

Grissmer, D. W., Subotnik, R. F., & Orland, M. (2009). *A guide to incorporating multiple methods in randomized controlled trials to assess intervention effects*. Washington, DC: American Psychological Association.

Grund, S., Lüdtke, O., & Robitzsch, A. (2016). *Multiple imputation of multilevel missing data: An introduction to the R Package pan*. SAGE Open, 6(4), 2158244016668220.

Hanushek, E. A. (1999). Some findings from an independent investigation of the Tennessee STAR experiment and from other investigations of class size effects. *Educational Evaluation and Policy Analysis*, *21*(2), 143-163.

Hart Research Associates. (2013). It takes more than a major: Employer priorities for college learning and student success. *Liberal Education.*, *99*.

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

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

Hedges, L. V., & Hedberg, E. C. (2013). Intraclass correlations and covariate outcome correlations for planning two-and three-level cluster-randomized experiments in education. *Evaluation Review*, *37*(6), 445-489.

Hedges, L. V., & Rhoads, C. (2010). *Statistical power analysis in education research. NCSER 2010-3006*. National Center for Special Education Research.

Hu, S., & Kuh, G. D. (2003). Maximizing what students get out of college: Testing a learning productivity model. *Journal of College Student Development*, *44*(2), 185-203.

Huber, C. R., & Kuncel, N. R. (2016). Does college teach critical thinking? A meta-analysis. *Review of Educational Research*, 86(2), 431-468.

IES Funding (2017, August 09). Retrieved from http://ies.ed.gov/funding/grantsearch

IES NCEE Programs (2017, August 09). Retrieved from https://ies.ed.gov/ncee/projects/ evaluation/evaluations_filter.asp

IES NCER Programs (2017, August 09). Retrieved from https://ies.ed.gov/ncer/projects/ program.asp?ProgID=15

IES NCEE Evaluation TA. (2017, August 09) Retrieved from https://ies.ed.gov/ncee/projects/ evaluationTA.asp.

Institute of Education Sciences (ED). (2003). *Identifying and implementing educational practices supported by rigorous evidence: A user friendly guide*. Washington, DC.: ERIC Clearinghouse.

Ishiyama, J. (2002). Does early participation in undergraduate research benefit social science and humanities students? *College Student Journal*, *36*, 380-386.

Jacob, R., Zhu, P., & Bloom, H. S. (2010). New empirical evidence for the design of group in education. *Journal of Research on Educational Effectiveness*, *3*(2), 157-198.

Kane, T. J. (2004, January 16). *The impact of after-school programs: Interpreting the results of four recent evaluations*. William T. Grant Foundation Working Paper. Retrieved July 16, 2008, from http://www.wtgrantfoundation.org/usr_doc/After-school_paper. pdf

Kelcey, B., Spybrook, J., Phelps, G., Jones, N., & Zhang, J. (2017). Designing large-scale multisite and cluster-randomized studies of professional development. *The Journal of Experimental Education*, *85*(3), 389-410.

Kim, M. M. (2001). Institutional effectiveness of women-only colleges: Cultivating students' desire to influence social conditions. *The Journal of Higher Education*, *72*(3), 287-321.

Kim, Y. K., & Sax, L. J. (2009). Student–faculty interaction in research universities: Differences by student gender, race, social class, and first-generation status. *Research in Higher Education*, *50*, 437-459.

Kim, Y. K., & Sax, L. J. (2011). Are the effects of student–faculty interaction dependent on academic major? An examination using multilevel modeling. *Research in Higher Education*, *52*, 589-615.

Kinzie, J., Thomas, A. D., Palmer, M. M., Umbach, P. D., & Kuh, G. D. (2007). Women students at coeducational and women's colleges: How do their experiences compare? *Journal of College Student Development*, *48*(2), 145-165.

Kitchener, K., Wood, P., & Jensen, L. (2000). Promoting epistemic cognition and complex judgment in college students. In annual meeting of the American Psychological Association, Washington, DC.

Klein, S., Freedman, D., Shavelson, R., & Bolus, R. (2008). Assessing school effectiveness. *Evaluation Review*, *32*(6), 511-525.

Klein, S. P., Kuh, G., Chun, M., Hamilton, L., & Shavelson, R. (2005). An approach to measuring cognitive outcomes across higher education institutions. *Research in Higher Education*, *46*(3), 251-276.

Klein, S. C., Liu, O. L. E., Sconing, J. A., Bolus, R. C., Bridgeman, B. E., Kugelmass, H. C., & Steedle, J. C. (2009). *Test Validity Study (TVS) Report*.

Klein, S., Steedle, J., & Kugelmass, H. (2010). *The Lumina longitudinal study: Summary of procedures and findings*. Unpublished manuscript.

Kobrin, J. L., Patterson, B. F., Shaw, E. J., Mattern, K. D., & Barbuti, S. M. (2008). *Validity of the SAT® for Predicting First-Year College Grade Point Average*. Research Report No. 2008-5. College Board.

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

Kugelmass, H., & Ready, D. D. (2011). Racial/ethnic disparities in collegiate cognitive gains: A multilevel analysis of institutional influences on learning and its equitable distribution. *Research in Higher Education*, *52*(4), 323-348.

Kuncel, N. R., Credé, M., Thomas, L. L., Klieger, D. M., Seiler, S. N., & Woo, S. E. (2005). A meta-analysis of the validity of the Pharmacy College Admission Test (PCAT) and grade predictors of pharmacy student performance. *American Journal of Pharmaceutical Education*, *69*(3), 51.

Lagemann, E. C. (1997). Contested terrain: A history of education research in the United States, 1890–1990. *Educational Researcher*, *26*(9), 5-17.

Lather, P., & Moss, P. A. (2005). Introduction: Implications of the scientific research in education report for qualitative inquiry. *Teachers College Record*.

Lee, W., Chiang, C. H., Liao, I. C., Lee, M. L., Chen, S. L., & Liang, T. (2013). The longitudinal effect of concept map teaching on critical thinking of nursing students. *Nurse Education Today*, *33*(10), 1219-1223.

Lipsey, M., Bloom, H., Hill, C., & Black, A. (2007, February 6). How big is big enough? Achievement effect sizes in education. Presented at University of Pennsylvania Graduate School of Education.

Little, R. J. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American statistical Association*, *83*(404), 1198-1202.

Liu, O. L. (2011). Value-added assessment in higher education: A comparison of two methods. *Higher Education*, *61*(4), 445-461.

Mccann, C., A., Laitinen, & Feldman, A. (January, 2018). Putting the experiment back in the experimental site initiative. Retrieved from https://na-roduction.s3.amazonaws.com/documents/Putting-Experiment-Back-Experimental-Sites-Initiative.pdf

McCormick, A. C., Pike, G. R., Kuh, G. D., & Chen, P. S. D. (2009). Comparing the utility of the 2000 and 2005 Carnegie classification systems in research on students' college experiences and outcomes. *Research in Higher Education*, *50*(2), 144-167.

Moberg, J., & Kramer, M. (2015). A brief history of the cluster randomized trial design. *Journal of the Royal Society of Medicine*, *108*(5), 192-198.

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

Mosteller, F., Light, R., & Sachs, J. (1996). Sustained inquiry in education: Lessons from skill grouping and class size. *Harvard Educational Review*, *66*(4), 797-843.

Murray, D. M. (1998). Design and analysis of group. New York: Oxford University Press USA.

populations. *Journal of Educational and Behavioral Statistics*, 25(3), 271-284.

Myers, D., Olsen, R., Seftor, N., Young, J., & Tuttle, C. (2004). *The Impacts of Regular Upward Bound: Results from the Third Follow-Up Data Collection*. MPR Reference No. 8464-600. Mathematica Policy Research, Inc.

Niehaus, E., Campbell, C. M., & Inkelas, K. K. (2014). HLM behind the curtain: Unveiling decisions behind the use and interpretation of HLM in higher education research. *Research in Higher Education*, *55*(1), 101-122.

Niu, L., Behar-Horenstein, L. S., & Garvan, C. W. (2013). Do instructional interventions influence college students' critical thinking skills? A meta-analysis. *Educational Research Review*, *9*, 114-128.

Norman, G. R., & Schmidt, H. G. (2000). Effectiveness of problem-based learning curricula: Theory, practice and paper darts. *Medical Education*, *34*(9), 721-728.

O'Connell, A. A., & Reed, S. J. (2012). Hierarchical data structures, institutional research, and multilevel modeling. *New Directions for Institutional Research*, *2012*(154), 5-22.

Orr, L. L. (1999). Social experiments. *Evaluating public programs with experimental methods*.

Ortiz, C. M. A. (2007). *Does philosophy improve critical thinking skills?* University of Melbourne, Department of Philosophy.

Pace, C. R. (1990). *The undergraduates: A report of their activities and progress in college in the 1980s.* Los Angeles: Center for the Study of Evaluation, Graduate School of Education London: Sage.

Pascarella, E. T. (1985). The influence of on-campus living versus commuting to college on intellectual and interpersonal self-concept. *Journal of College Student Personnel*.

Pascarella, E. T., Blaich, C., Martin, G. L., & Hanson, J. M. (2011). How robust are the findings of academically adrift? *Change: The Magazine of Higher Learning*, *43*(3), 20-24.

Pascarella, E. T., & Terenzini, P. T. (1991). *How College Affects Students*. San Francisco, CA: Jossey Bass.

Pascarella, E. T., & Terenzini, P. T. (2005). *How College Affects Students* (Vol. 2). K. A. Feldman (Ed.). San Francisco, CA: Jossey-Bass.

Peugh, J. L., & Enders, C. K. (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of Educational Research*, *74*(4), 525-556.

Porter, A., McMaken, J., Hwang, J., & Yang, R. (2011). Common core standards the new US intended curriculum. *Educational Researcher*, *40*(3), 103-116.

Quartagno, M., & Carpenter, J. (2016). jomo: A package for multilevel joint modelling multiple imputation. R package version, 2-2.

Quint, Jennet (2011). *Research advances: Using luster random assignment*. Retrieved from: http://www.mdrc.org/publication/research-advances-using-cluster-random-assignment

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

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1). Sage.

Raudenbush, S. W., & Liu, X. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological Methods*, *5*(2), 199.

Raudenbush, S. W., & Liu, X. F. (2001). Effects of study duration, frequency of observation, and sample size on power in studies of group differences in polynomial change. *Psychological Methods*, *6*(4), 387.

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.

Richburg-Hayes, L. (2015). Reauthorizing the Higher Education Act: Opportunities to Improve Student Success. Additional Submitted Testimony from Lashawn Richburg-Hayes, MDRC, to the Senate Committee on Health, Education, Labor, and Pensions. MDRC.

Ross, S. M., Morrison, G. R., & Lowther, D. L. (2005). Using experimental methods in higher education research. *Journal of Computing in Higher Education*, *16*(2), 39-64.

Rothstein, J. (2005, June). SAT scores, high schools, and collegiate performance predictions. In annual meeting of the National Council on Measurement in Education, Montreal, CA.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*(2), 147.

Schafer, J. L., & Yucel, R. M. (2002). Computational strategies for multivariate linear mixed-effects models with missing values. *Journal of Computational and Graphical Statistics*, *11*(2), 437-457.

Schanzenbach, D. W. (2006). What have researchers learned from Project STAR?. *Brookings Papers on Education Policy*, *9*, 205-228.

Schmidt, H. G. (1983). Problem-based learning: Rationale and description. *Medical Education*, *17*(1), 11-16.

Schochet, P. Z. (2008). *Statistical power for random assignment evaluations of education programs*.

Schultz, J. L., & Mueller, D. (2006). *Effectiveness of programs to improve postsecondary* *education enrollment and success of underrepresented youth: A literature review*. St. Paul, MN: Wilder Research.

Schwartz, J., & Szabo, Z. (2011). Targeted instruction for preservice teachers: Developing higher order thinking skills with online discussions. *International Journal of Education and Psychology in the Community, 1*(1), 32-51.

Scrivener, S., & Coghlan, E. (2011). *Opening Doors to Student Success: A Synthesis of Findings from an Evaluation at Six Community Colleges*. New York: MDRC.

Scrivener, S., & Weiss, M. J. (2009). *More guidance, better results? Three-year effects of an enhanced student services program at two community colleges*. New York, NY: MDRC.

Scrivener, S., Weiss, M. J., Ratledge, A., Rudd, T., Sommo, C., & Fresques, H. (2015). Doubling graduation rates: Three-year effects of CUNY's Accelerated Study in Associate Programs (ASAP) for developmental education students. *Scrivener, Susan, Michael J. Weiss, Alyssa Ratledge, Timothy Rudd, Colleen Sommo, and Hannah Fresques, Doubling Graduation Rates: Three-Year Effects of CUNY's Accelerated Study in Associate Programs (ASAP) for Developmental Education Students. New York: MDRC*.

Shavelson, R. J., and Huang, L. (2003). Responding responsibly to the frenzy to assess learning in higher education. *Change: The magazine of higher education learning, 35*(1): 10–19.

Silva, E. (2008). Measuring Skills for the 21st Century. Education Sector Reports. *Education Sector*.

Snijders, T., & Bosker, R. (1999). *Multilevel analysis: An introduction to basic and applied multilevel analysis*.

Sohn, K. (2016). Improved deep metric learning with multi-class n-pair loss objective. In *Advances in Neural Information Processing Systems* (pp. 1857-1865).

Song, M., & Herman, R. (2010). Critical issues and common pitfalls in designing and conducting impact studies in education: Lessons learned from the What Works Clearinghouse (Phase I). *Educational Evaluation and Policy Analysis*, *32*(3), 351-371.

Spellings, M. (2006). *A test of leadership: Charting the future of US higher education*. Washington, DC: U.S. Department of Education.

Spybrook, J. (2008). Power, sample size, and design. *Multilevel modeling of educational data*, 273-311.

Spybrook, J. (2013). Introduction to special issue on design parameters for cluster randomized trials in education. *Evaluation Review, 37*(6), 435-444.

Spybrook, J. (2014). Detecting intervention effects across context: An examination of the precision of cluster. *The Journal of Experimental Education*, *82*(3), 334-357.

Spybrook, J., Bloom, H., Congdon, R., Hill, C., Martinez, A., Raudenbush, S., & TO, A. (2011). *Optimal design plus empirical evidence: Documentation for the "Optimal Design" software*. William T. Grant Foundation. Retrieved on November 5, 2012.

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

Stark, E. (2012). Enhancing and assessing critical thinking in a psychological research methods course. *Teaching of Psychology*, *39*(2), 107-112.

Steedle, J. T. (2012). Selecting value-added models for postsecondary institutional assessment. *Assessment & Evaluation in Higher Education, 37*(6), 637-652.

Stieff, M., Superfine, A., Yin, Y. (September, 2017). *Efficacy of the connected chemistry curriculum*. Retrieved from https://ies.ed.gov/funding/grantsearch/details.asp?ID=1941

Tabachnick, B. G., & Fidell, L. S. (2001). *Using Multivariate Statistics* (4th ed.). Needham Heights, MA: Allyn and Bacon.

The New Commission on the Skills of the American Workforce. (2006). *Tough Choices or Tough Times*. Washington, DC: National Center on Education and the Economy.

The Secretary's Commission on Achieving Necessary Skills. (1991). *What work requires of schools: A SCANS Report for America 2000*. Washington, DC: U.S. Department of Labor.

Tiruneh, D. T., Verburgh, A., & Elen, J. (2014). Effectiveness of critical thinking instruction in higher education: A systematic review of intervention studies. *Higher Education Studies, 4*(1), 1-17.

Towne, L., & Shavelson, R. J. (2002). *Scientific research in education*. Washington, DC: National Research Council.

Tremblay, K., Lalancette, D., & Roseveare, D. (2012). Assessment of higher education learning outcomes: Feasibility study report, volume 1 design and implementation.

Van Buuren, S. (2011). Multiple imputation of multilevel data. *Handbook of advanced multilevel analysis*, *10*, 173-196.

Van Gelder, T. (2005). Teaching critical thinking. *College teaching*, *45*(1), 1-6.

Variance Almanac of Academic Achievement. (2017, October 20). Retrieved from https://arc.uchicago.edu/reese/variance-almanac-academic-achievement

Visher, M., Butcher, K. F., & Cerna, O. S. (2011). *Guiding Math Students to Campus Services: An Impact Evaluation of the Beacon Program at South Texas College*. Evanston, IL: Society for Research on Educational Effectiveness. (ERIC Document Reproduction Service No. ED 517 927).

Visher, M. G., Weiss, M. J., Weissman, E., Rudd, T., & Wathington, H. D. (2012). *The Effects of Learning Communities for Students in Developmental Education: A Synthesis of Findings from Six Community Colleges*. National Center for Postsecondary Research.

Volkwein, J. F., King, M. C., & Terenzini, P. T. (1986). Student-faculty relationships and intellectual growth among transfer students. *The Journal of Higher Education*, *57*(4), 413-430.

Voluntary System of Accountability (2017). Retrieved from http://www.voluntarysystem.org/

Watson, G. (1980). *Watson-Glaser critical thinking appraisal*. San Antonio, TX: Psychological Corporation.

Weidman, J. (1989). Undergraduate socialization: A conceptual approach. *Higher education: Handbook of theory and research*, *5*(2), 289-322.

Weiss, M. J., Mayer, A. K., Cullinan, D., Ratledge, A., Sommo, C., & Diamond, J. (2015). A random assignment evaluation of learning communities at Kingsborough Community College—Seven years later. *Journal of Research on Educational Effectiveness*, *8*(2), 189-217.

Westine, C., Spybrook, J., & Taylor, J. (2013). An empirical investigation of design parameters for planning cluster of science achievement. *Evaluation Review*, *37*(6), 490-519.

What Works Clearinghouse (March, 2018). *WWC Procedures and Standards Handbook, Version 4.0*. Retrieved from https://ies.ed.gov/ncee/wwc/Docs/referenceresources/ wwc_procedures_handbook_v4.pdf

Wheeler, L. A., & Collins, S. K. (2003). The influence of concept mapping on critical thinking in baccalaureate nursing students. *Journal of professional Nursing*, *19*(6), 339-346.

Whitehurst, G. (2002). Charting a new course for the U.S. Office of Educational Research and

Improvement. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Whitt, E., Edison, M., Pascarella, E., Nora, A., & Terenzini, P. (1999). Interactions with peers and objective and self-reported cognitive outcomes across three years of college. *Journal of College Student Development*, *40*, 61-78.

Wolf, A., Frye, M., Goodson, B., Price, C., & Boulay, B. (2016). Example Evaluation Plan for a Cluster Randomized Controlled Trial.

Woo, S. E. (2005). A meta-analysis of the validity of the Pharmacy College Admission Test (PCAT) and grade predictors of pharmacy student performance. *American Journal of Pharmaceutical Education*, *69*(3), 51.

Xu, Z., & Nichols, A. (2010). *New estimates of design parameters for clustered randomization studies: Findings from North. Carolina and Florida*. Washington, DC: Urban Institute, National Center for Analysis of Longitudinal Data in Education Research.

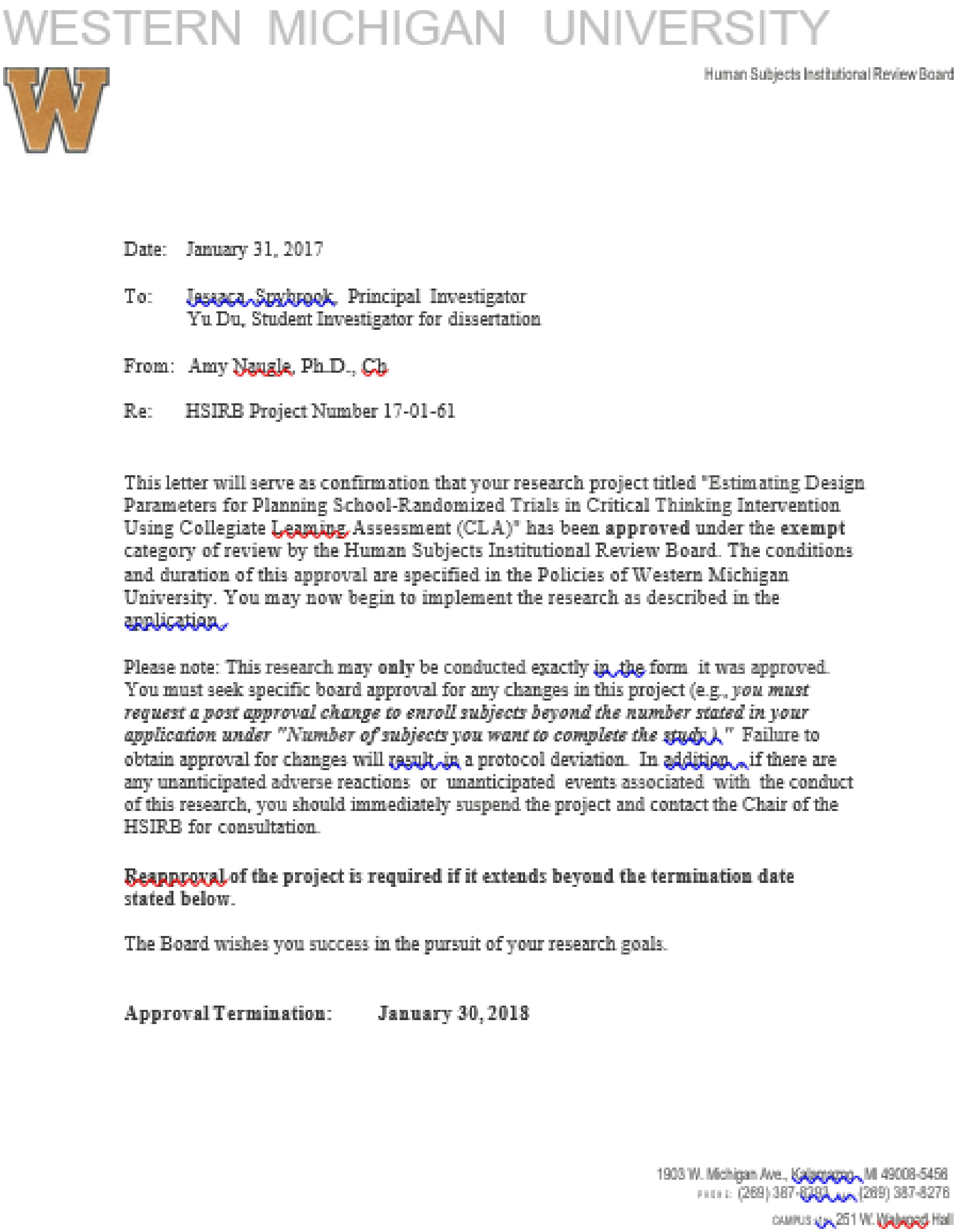
Yang, Y. T. C., & Chou, H. A. (2008). Beyond critical thinking skills: Investigating the relationship between critical thinking skills and dispositions through different online instructional strategies. *British Journal of Educational Technology, 39*(4), 666-684.

Zahner, D., Steedle, J. (2015). Comparing longitudinal and cross-sectional school effect estimates in postsecondary education. Retrieved from http://cae.org/images/uploads/pdf/Comparing_Longitudinal_and_Cross-Sectional_School_Effect_Estimates.pdf

Zhao, C. M., & Kuh, G. D. (2004). Adding value: Learning communities and student engagement. *Research in higher education*, *45*(2), 115-138.

Zhu, P., Jacob, R., Bloom, H. S., & 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.

Appendix



WESTERN MICHIGAN UNIVERSITY

Human Subjects Institutional Review Board

Date: January 31, 2017

To: Jessaca Spybrook, Principal Investigator
Yu Du, Student Investigator for dissertation

From: Amy Naugle, Ph.D., Ch

Re: HSIRB Project Number 17-01-61

This letter will serve as confirmation that your research project titled "Estimating Design Parameters for Planning School-Randomized Trials in Critical Thinking Intervention Using Collegiate Learning Assessment (CLA)" 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: January 30, 2018

1903 W. Michigan Ave., Kalamazoo, MI 49008-5456
PHONE: (269) 387-8293 FAX: (269) 387-8276
CAMPUS SITE: 251 W. Walwood Hall