8-1-2013

Assessing sensitivity of Early Head Start study findings to manipulated randomization threats

Sheridan Green

UNIVERSITY OF NORTHERN COLORADO

Greeley, Colorado

The Graduate School

ASSESSING SENSITIVITY OF EARLY HEAD START STUDY FINDINGS TO MANIPULATED RANDOMIZATION THREATS

A Dissertation Submitted in Partial Fulfillment
of the Requirements for the Degree of
Doctor of Philosophy

Sheridan Green

College of Education and Behavior Sciences
Department of Applied Statistics and Research Methods

August 2013

This Dissertation by: Sheridan Green

Entitled: *Assessing Sensitivity of Early Head Start Study Findings to Manipulated Randomization Threats*

has been approved as meeting the requirement for the Degree of Doctor of Philosophy in College of Education and Behavior Sciences in Department of Applied Statistics and Research Methods

Accepted by the Doctoral Committee

____
Susan R. Hutchinson, Ph.D., Research Advisor

____
Lisa Rue, Ph.D., Committee Member

____
Antonio Olmos, Ph.D., Committee Member

____
Amanda J. Moreno, Ph.D., Committee Member

____
Linda L. Black, Ed.D., LPC, Faculty Representative

Date of Dissertation Defense_______________________________

Accepted by the Graduate School

____
Linda L. Black, Ed.D., LPC
Acting Dean of the Graduate School and International Admissions

ABSTRACT

Green, Sheridan. *Assessing Sensitivity of Early Head Start Study Findings to Manipulated Randomization Threats.* Published Doctor of Philosophy dissertation, University of Northern Colorado, 2013.

Increasing demands for design rigor and an emphasis on evidence-based practice on a national level indicated a need for further guidance related to successful implementation of randomized studies in education. Rigorous and meaningful experimental research and its conclusions help establish a valid theoretical and evidence base for educational interventions and curricula. The validity of findings derived from an experimental design largely depends on the quality of the randomization and the study implementation. This study's purpose was to systematically examine how the magnitude and type of typical randomization problems affected study results. I used secondary data from a randomized national study, the Early Head Start Research and Evaluation Project, to examine the manipulated effects of threats to randomization on selected child developmental outcomes. Data were exposed to 1 of 27 different threat conditions and were compared with randomized data using sensitivity analysis to assess effects on intervention-control balance on covariates and results bias introduced by the threat (Type I and II error, mean percent bias, and effect size differences). The conditions varied by overall sample size (small, medium, large), proportion of the sample disrupted (5%, 15%, 25%), and the type of disruption (allocation bias, noncompliance, differential attrition).

The effects of post hoc statistical adjustments, propensity score analysis, and analysis of covariance were also examined.

The introduction of imbalance and bias in baseline covariates generally led to bias in the results under threat conditions. The allocation bias scenario was most affected by imbalance under the threat condition, although a high level of imbalance was also introduced in the noncompliance scenarios and a moderate amount in the differential attrition conditions. Baseline imbalance and bias were greatest for the large samples and for the samples that were threatened at the 25% threat level. As expected, the greater the proportion of the sample affected by the threat scenario, the greater the likelihood of baseline imbalance, bias, and biased results.

The threat scenario under which the outcomes results were most sensitive was allocation bias, matching the larger baseline imbalance and bias introduced. Examination by sample size indicated a relatively high rate of Type I error was found for the small samples, while the highest rates of Type I and Type II error were among the large samples. Overall bias was highest among the small samples; 35% of the tests were biased either in terms of the significance test, the mean effect size difference, or mean percent bias. The samples that were affected by the largest proportion of threat were the most sensitive to the disruption, again matching the levels of introduced baseline imbalance and bias. For this group, the high mean percent bias (14.3%) was notably higher than the 15% and 5% threat levels. Overall, the adjustment techniques introduced more bias than they corrected.

The well-respected randomized design is susceptible to any number of design threats which, depending on the circumstances, might bias effect estimates of interest.

Implications for researchers, regardless of study sample size, include measuring sufficient baseline covariates to conduct balance checks, preventing design threats by closely monitoring research practices, and generally using various means, such as literature review and replication, to cross-check findings. Statistical adjustment methods appropriate to the threat type are warranted when bias is likely to be present. The reliance on randomization to prevent all internal validity problems should be in direct proportion to efforts taken to maintain the design's integrity.

ACKNOWLEDGEMENTS

The findings reported here are based on research conducted as part of the national Early Head Start Research and Evaluation Project funded by the Administration for Children and Families (ACF), U.S. Department of Health and Human Services under contract to Mathematica Policy Research, Princeton, New Jerey, and Columbia University’s National Center for Children and Families, Teachers College, in conjunction with the Early Head Start Research Consortium. The Consortium consists of representatives from 17 programs participating in the evaluation, 15 local research teams, the evaluation contractors, and ACF.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AIC	Akaike Info Criteria
ANCOVA	Analysis of covariance
ANOVA	Analysis of variance
BIC	Bayesian Information Criteria
BID-II	Bayley Scales of Infant Development--Second Edition
CACE	Complier Average Causal Effect
CBCL	Child Behavior Checklist
EHS	Early Head Start
EHSRE	Early Head Start Research and Evaluation Project
ES	Effect Size
MDI	Mental Developmental Index
MNAR	Missing-Not-At-Random
NCLB	No Child Left Behind
NICHD	National Institute of Child Health and Human Development
OBS	Original Base Set
PPVT-III	Peabody Picture Vocabulary Test-III
PSA	Propensity Score Analysis

CHAPTER I

INTRODUCTION AND BACKGROUND

Increasing focus on reform and accountability over the past decade has driven numerous political, methodological, and statistical advances to support more rigorous and effective research and programs in education in the United States (Burtless, 2001; Hostetler, 2005; Institute of Education Sciences, U.S. Department of Education [IES], 2008; National Research Council [NRC], 2002; No Child Left Behind, 2002). The call for evidence-based programs in education founded on scientifically-based research (i.e., grounded in rigorous principles and methods; NRC, 2002) has grown ever louder since 2001, advocating the use of randomized experimental designs, comparing outcomes for intervention and control groups, and promoting the conduct of more rigorous quasi-experiments.

Randomized experimental studies are known by a variety of names: randomized controlled trials, randomized field experiments, and, simply, experiments. Experiments employ the technique of randomization, also called random assignment, and can be classified into different design types (Campbell & Stanley, 1963). Random assignment is defined as the selection method in an experiment by which an experimental subject or participant has an equal probability of being chosen for one of *n* available interventions, treatments, or conditions (Fisher, 1935; Hall, 2002; Levi, 1983). Gueron (2001) described random assignment another way, explaining that random assignment uses a "lottery-like process" to allocate participants to two or more groups to be compared to determine the intervention's "net impact," with the net impact referring to the outcomes achieved that would not have resulted without the intervention (p. 18).

Chapter II of this work more thoroughly describes the purpose and advantages of random assignment; however, in brief, the historic and current advocacy for random assignment is aimed mainly at improving researchers' ability to determine the causal effects of interventions, e.g., educational programs and curricula implementation. The goal is to understand whether (and to what extent) educational interventions have an impact on the outcomes under study while controlling for other potential influences called confounds or extraneous variables. Gall, Gall, and Borg (2003) described extraneous variables as any variable or aspect of the situation other than the intervention that might affect the study outcome if not controlled in the context of the experiment (p. 368).

Random assignment is valued for many reasons, one being its ability to reduce selection bias, i.e., bias associated with systematic differences between intervention and control groups. Bias refers to any systematic error encountered in a study. Random assignment increases the ability to obtain unbiased causal estimates (i.e., results) because at pretest, the intervention and control groups are considered equivalent on relevant covariates (Steiner, Cook, Shadish, & Clark, 2010). Other benefits include the ability to minimize or control most threats to internal validity (Berger, 2005; Campbell & Stanley, 1963; Gall et al., 2003). The fear with poorly executed research designs that are intended to describe causal relationships is that differentiation between spurious, indirect, and causal associations becomes very difficult (Grimes & Schulz, 2002).

The following provocative statement made by key leaders in educational and cross-disciplinary research (Cook & Shadish, 2011) reflects both the hope and frustration of the field: “the state of causal research in education is *sorry, but improving*” (Northwestern University, 2011, p. 1). Indeed, relatively recent innovations have contributed to considerable improvements in quasi-experiments of education interventions. For example, increased use of propensity score analysis and regression discontinuity design improves study rigor by helping reduce group assignment selection bias and has arguably improved researchers’ ability to make causal claims. To briefly explain, propensity score methods involve the statistical computation of a score that can be used to equate intervention and control groups on those covariates through matching, stratification, weighting, regression, or ANCOVA (Analysis of Covariance; Clark, 2008; Shadish & Steiner, 2010). This is generally an unnecessary process in randomization because by definition the groups are considered equivalent by chance. Regression discontinuity design is a method that compares intervention and control groups’ scores within a relatively narrow band around a cut-off point or threshold. Regression discontinuity design is considered the next-best method to RA in reducing selection bias (Schochet et al., 2010).

Regardless of these advances, no matter what field or sector, randomizing study participants to intervention or control groups is still considered the most effective way to obtain unbiased average intervention effects (Fisher, 1925, 1935; Gall et al., 2003; Hall, 2007; National Research Council, 2002; Steiner, Cook, Shadish, & Clark, 2010). These benefits remain the justification for continued advocacy for using randomized designs in educational intervention research. Evidence suggests, however, that randomized studies have not been frequently conducted in education research compared with other fields (Boruch, DeMoya, & Snyder, 2001; Cook, 2002, 2003; Cook & Shadish, 1994), although their numbers are slowly increasing. Indeed, U.S. research in education is barely 100 years old (National Research Council, 2002).

Despite the potential benefits of using an experimental design, unsuccessful, inappropriate, or incorrect implementation of this design could weaken or eliminate the bias-reducing advantages of random assignment. It is within this context that the current research was conducted. Threats to successful randomization are numerous and comprised the primary focus of this study. The following paragraphs of this chapter describe the problems researchers face with regard to randomization threats, provide a rationale for the study, and briefly explain the context of the study. This chapter also contains a description of the study's purpose, research questions, limitations, and implications.

Problem Statement

Researchers generally agree that random assignment, when properly implemented, remains the best way to ensure unbiased study conclusions regarding intervention effectiveness and to enhance potential for inferring causality with respect to intervention or treatment effects (Fisher, 1925, 1935; National Research Council, 2002; Papineau, 1994). Berger (2005) admitted, however, that there is a continuum of quality among randomized studies that often goes unacknowledged (i.e., not all randomized studies are alike). He argued that when the term *random assignment* is used, the internal validity of the study findings is assumed, when in fact it is not guaranteed. Perfectly implemented randomized intervention studies in the social sciences are incredibly rare

(Dunn, Maracy, & Tomenson, 2005). While successful implementation of random assignment is powerful and effective, *unsuccessful* implementation or post-randomization problems clearly pose serious threats to the design's inherent benefits. Researchers' ability to interpret study findings and infer causality diminishes when random assignment is compromised. The added problem is that the degree to which, and under what conditions, certain random assignment problems do affect the results is unknown, particularly in educational intervention studies. A limited number of causal investigations using randomized designs exists in education (Boruch et al., 2001; Maxwell, 2004; Vinovskis, 2002). Thus, more than a little potential for design implementation problems exists from a lack of cumulative experience. A possible explanation for the lack of experimental studies in education has been inconsistent political support and funding for evaluation in education over the last century (see Vinovskis, 2002, for a brief history of development and evaluation at the U.S. Department of Education).

The assumption that the use of a randomized design just "takes care" of design problems is to a large degree true when it is implemented appropriately. In fact, implementing random assignment seems relatively simple compared with the challenging statistical techniques and adjustments required to produce good quasi-experiments. However, random assignment studies are commonly hijacked by problems such as biased allocation, noncompliance, and differential attrition (Downs, Tucker, Christ-Schmidt, & Wittes, 2010), making them more vulnerable than generally recognized.

Randomization disruption, contamination, or failure might occur via various problems associated with the randomization process itself. For example, flawed randomization might occur in using methods inappropriate for the design, sample characteristics, or desired level of analysis. Randomization failure is likely when there is biased (flawed) allocation (e.g., when the assignment to groups intended to be random is not). Alternatively, disruption might be caused by post-randomization problems like intervention noncompliance or differential attrition. To define, noncompliance occurs when research participants fail to follow their allocated intervention assignments (Chen et al., 2009; Dunn et al., 2005; Frangakis & Rubin, 1999). For instance, a control participant might unexpectedly receive intervention or a participant allocated to intervention might not attend scheduled intervention sessions. This is a problem because comparison of the intervention effects by group might become attenuated. Differential attrition is essentially a missing data/nonresponse problem and it refers to systematic differences between the intervention and control groups that might influence the study results. Differential attrition is considered to be one of the greatest threats to internal validity (Foster & Bickman, 1996) and is common in longitudinal, randomized experiments (Shadish, Hu, Glaser, Kownacki, & Wong, 1998). Both small and large scale studies are at risk of encountering threats to randomization of one kind or another. These problems also lead to biased and invalid effect estimates for the intervention study. The use of a randomized design is specifically intended to ward off selection bias; however, it cannot prevent measurement bias (i.e., how the outcome was measured) and intervention bias (e.g., noncompliance). On the other hand, the latter bias types can disrupt the advantages gained by randomizing. Incorrect findings can foster misconceptions in interpretations and conclusions that ultimately mislead the field. As an added insult, they can destroy the utility of said findings in supporting the development of effective evidence-based policies and interventions.

The climate, culture, and context of educational interventions make for a unique combination and possible increased likelihood of randomization difficulties (De Anda, 2007; Falaye, 2009; National Research Council, 2002; Ong-Dean, Hofstetter, & Strick, 2010). For example, conducting randomized studies in education is still not particularly well supported by families, school and program staff, and local school administration (Gueron, 2001). This lack of support comes from many factors including a widespread perception of unfairness or that it is unethical to withhold program or educational services from one group but not another (Dunford, 1990; Gueron, 2001). Given this belief, randomization processes might be thwarted in some cases.

Clearly, advancement and innovation have contributed to quality improvements in methods and overall rigor in design; however, warnings about inferring causation in quasi-experimental designs and poorly implemented experiments remain. Intentional efforts to improve the state of causal research in education are needed to enhance the quality of research findings and the validity of conclusions. Threats to randomization are in direct opposition to this effort.

Context for the Current Study

A number of theoretical, political, and cultural contexts influenced the conduct and exigency of this study. The framework of causality is well matched with the relatively recent emphasis on ensuring educational programming is developed on scientifically-based research. The focus of this study was to understand the impact of design threats on the ability to ultimately make valid causal claims. It is essential to know what interventions and programs work. The evidence base of what works (and what does not work) in educational programs has been growing but is still inadequate. Likewise, improved methodological knowledge related to the implementation of randomized designs is necessary (Cook, 2002; Gueron, 2001; Imai, Keele, Tingley, & Yamamoto, 2011; Shadish & Steiner, 2010; Towne & Hilton, 2003). Chapter II provides a more thorough discussion of causality and what is known as the counterfactual framework.

Call for Increased Rigor

Perhaps in reaction to their scarcity, a broad increase in the number of randomized studies in education is expected due to recent legislative changes and political shifts. Legislation guides the mission and functions of the U.S. Department of Education and its stance toward scientific research. The Education Sciences Reform Act of 2002 specified the establishment of the U.S. Department of Education, Institute of Education Sciences (IES), and its various centers for research, evaluation, and statistics. Within the Act, standards for the conduct and evaluation of research are outlined, clearly setting the expectations for adherence to and development of scientifically-based research standards (Education Sciences Reform Act, 2002). Institute of Education Sciences'(2011) subsequent evolution over the last decade is highlighted by the development of research priorities emphasizing increased rigor and relevance, which includes heightened provision of methodological and analytic support. It is within this national context of improving research in education that the current study is housed--one of increasing demands for research rigor and causal claims.

No Child Left Behind Legislation

No Child Left Behind **(**NCLB; 2002) provided another contextual influence on the current study. The climate fostered by NCLB and other current accountability frameworks is characterized by a top down, audit-like approach to accountability, assessment, and evaluation. Evaluators focused on educational evaluation theory face many challenges in the accountability evaluation legislated by No Child Left Behind (Ryan, 2004, p. 444). This presents difficulties with regard to teacher support of evaluation at the classroom and school levels. Within a culture of fear of losing one's job, teachers are likely to view evaluation with much more suspicion and reluctance (Gueron, 2001). With the added obligations associated with the requirements of a randomized design, teachers are even more likely to withhold support for research and evaluation in their classrooms. In many ways, the climate set by NCLB detracts from the goals to grow a deeper understanding from research about what works in educating young children; yet this climate does demand that more dedicated attention and energy go toward examining the state of affairs in policy and research in education.

What is not clearly articulated in the research literature or written governmental guidance is how standards of methodological rigor must be employed and maintained within the context of a randomized design. In addition, no formal guidance is available on the magnitude of detrimental effects (caused by certain types of randomization problems) on the accuracy of study results. With the understanding that no perfect implementation of randomization exists in research in education, it would certainly behoove researchers to be familiar with the comparative effects of small, medium, and large problems affecting randomization. So far the message is simply to use, whenever possible, randomization or an approximate, observational study alternative, meaning quasi-experiments carefully crafted to achieve accurate, unbiased study results.

Study Rationale

With the current climate of increasing demands for design rigor and an emphasis on evidence-based practice on a national level, it is apparent that further guidance is needed regarding successful implementation of randomized studies in education. Rigorous and meaningful experimental research and its conclusions help establish a valid theoretical and evidence base for educational interventions and curricula. The quality of findings that result from using an experimental design largely depends on the quality of RA implementation, the degree of the design's success in eliminating confounds and selection bias, and achieving *balance* across intervention and control groups. Balance in this sense means that intervention and control groups have equivalent distributions on available pre-intervention covariates (Hansen & Bowers, 2008; Steiner et al., 2010). As stated earlier, problems encountered in implementing randomization or in maintaining balance can lead to poor quality studies. Certainly one could argue that researchers are ethically obligated to understand, avoid, and combat the potential problems associated with the best-advocated, gold standard of research design. In the field of education, little comprehensive guidance on the differential effects of common randomization threats on study results exists, particularly as part of a strategic methodological comparison. Rather, in the rare instances when randomization problems are mentioned in the literature, they are typically an *a posteriori* explanation of studies-gone-wrong (Downs et al., 2010). Threats to randomization studies in education need to be examined because (a) they are relatively common; (b) conditions in the field of education pose a unique combination of threats to randomization; (c) the effects of randomization failure might lead to invalid study findings and interpretations; and (d) findings resulting from randomized studies in education have high stakes implications for policies and programming.

In the current study, I investigated different types and magnitudes of randomization threats in terms of their comparative effects on study results in the context of an early childhood intervention study. Most of the prior research in this area has been conducted using simulated data (Bruhn & McKenzie, 2009; Davies, Williams, & Yanchar, 2008; Manolov, Solanas, Bulté, & Onghena, 2010); thus, a further rationale for this study was the need to examine potential threats to randomization within a more complex real life situation with “real” data (see the limitations section below for a summary of the relative pros and cons of conducting this study using real versus simulated data). Understanding the effects of randomization disruption, particularly on real, rather than simulated, intervention data might support more vigilant research practices and increase awareness and prevention of potential problems. It was also intended that the study would address a critical question of “how much” disruption a study of this type could be tolerated before the assumption of strongly ignorable intervention assignment was violated, thus diminishing the bias-reducing advantages of the design by introducing confounds.

The rationale for this study centered on a need to make explicit the effects of common randomization threats to the accuracy and validity of study results. The intention was to explore whether the findings, obtained by analyzing data under a variety of manipulated randomization threat conditions, correctly matched the original randomized study findings. It was hoped that this study would contribute practical information in a realistic framework to the research improvement efforts so desired in the field of education. Falaye (2009) suggested there is a "need for researchers to update their skills in the use of randomized experiments through capacity building to ensure that findings from educational research and evaluation are creditable and defensible" (p. 21). Thus, gaining a better understanding of how common randomization threats in education studies affect the accuracy of study findings might help improve researchers' capacity to address problematic random assignment designs.

Purpose of the Study

The purpose of this study was to systematically examine how the magnitude and type of typical randomization problems affect study results. More specifically, I used secondary data from a randomized, national early childhood study, the Early Head Start Research and Evaluation Project (EHSRE), to examine the manipulated effects of threats to randomization on selected child developmental outcomes. In Phases I through III of the study, copies of foundation datasets were each exposed to 1 of 27 different threat conditions (using two methods of threat exposure each) and were compared with the EHSRE randomized datasets using sensitivity analysis to assess effects on the following: (a) intervention-control balance on covariates, (b) bias levels, (c) Type I and II error rates (depending on significance of findings on non-threatened samples), and (d) results on child outcome measures. The 27 conditions varied by overall sample size, proportion of the sample disrupted, and the type of disruption.

In Phase IV, I examined several statistical approaches to correcting or adjusting differences between the threatened and randomized data to reduce bias that might be introduced when randomization is compromised. In this phase, the tested variables included the types of corrective techniques employed and the dependent variables included the rates of successful correction to outcome and the percentage of bias reduction achieved by each method. The study was intended to increase knowledge of the potentially differential effects of randomization threats and the comparative effects of subsequent post-randomization statistical adjustment techniques commonly used in quasi-experimental and randomized designs.

Research Questions

Q1 What are the comparative effects of 27 randomization threat conditions on the randomization of the EHSRE data?

a. What evidence of intervention-control group covariate imbalance (i.e., baseline inequivalence between intervention and control groups) is revealed in each of 27 threat conditions?

b. What is the level of bias introduced by each of 27 threat conditions?

Q2 How sensitive are EHSRE study results to randomization threat conditions that include manipulations of threat type (i.e., biased allocation, noncompliance, and differential attrition), overall sample size, and proportion of sample exposed to the threat condition? Specifically,

a. To what extent do threatened and non-threatened samples differ on child outcome scores and observed effect sizes?

b. What is the rate of Type I and Type II error associated with the threatened samples compared with the associated non-threatened sample for each of the child outcome variables?

Q3 To what degree are corrective statistical methods effective in restoring or distorting results in the face of typical randomization threat conditions? Specifically,

a. What effect do corrective methods have on findings (means, significance, and effect sizes) from threatened samples?

b. Within threat types, what is the comparative effectiveness of two corrective techniques in reducing bias introduced under threat conditions?

Limitations and Scope

The scope of this investigation addressed three typical types of randomization disruption that could potentially occur. It is unknown the extent to which the study findings could generalize to what might be found across other randomization threat conditions. However, given that the set of underlying problems created by those threats (i.e., introduction of selection bias, missing data problems, and internal design validity problems) was similar across threat types, the chances for generalizability were good. In addition, since the study examined what occurred in a variety of sample size conditions, findings have a greater likelihood of applicability to both large, medium, and small studies.

Findings could pertain to studies of similar design, type, function, and size; this study might offer a good road map to design vigilance in the context of early childhood interventions. This study's findings were somewhat limited in terms of the correction methods employed. Replication of this study's findings would certainly be necessary for independent confirmation of findings. However, the correction methods I used were selected for their high frequency of use in the field to reduce selection bias and address missing data.

Perhaps the most important limitation of the study was that since a Monte Carlo simulation study was not conducted, the study did not achieve a simulated level of precise control over conditions. However, the trade-off was the results would be contextualized in a real-life intervention setting. Using these data provided a deliberate avenue for examining more realistic study situations that accurately reflected the complexity and quantity of independent and dependent variables common in large, randomized research. Because the intention of the study was to (a) present the effects of realistic, potential randomization disruption scenarios in an applied data setting; (b) understand the effects of said randomization disruption on complex study conclusions; and (c) offer realistic ameliorative strategies, it was decided that it would be more effectual to vary conditions within a real, randomized data set. Although a Monte Carlo simulation study would no doubt result in a greater degree of control with regard to the strict manipulation of levels of bias, the intent of this study was to examine these threats within the complexity of the existing biases of this large scale, randomized early childhood study.

Study Implications

The findings from this study have implications for the prevention, identification, and amelioration of randomization disruption in research studies in education. The results from this study were also used to develop a set of cautions and recommendations for researchers with respect to identifying, preventing, understanding, and correcting randomization problems under a variety of research and intervention conditions (see Discussion in Chapter V). It is intended that this information will enhance researchers' ability to obtain more accurate study findings about whether interventions work. Findings from well-run randomized studies in education will ultimately lead to better evidence about how to enhance children's achievement and well-being (Boruch et al., 2001). Ultimately, the consistent application of high quality research to practice in education is the key.

Summary

The current study was intended to investigate the differential effects of problems commonly occurring in studies using randomized designs. It used a secondary dataset from an experimental national early childhood intervention study. While experimental design is known for its many advantages in supporting causal claims with regard to the effectiveness of educational interventions, potential pitfalls within each of the phases of implementing such a design might dramatically and negatively affect the accuracy of study findings. Given the need and call for increased rigor in research and evaluation in education, supporting advances in what is known about potential challenges in RA implementation could not occur at a better time. It is hoped this study's findings provide a timely and relevant addition to the educational research literature.

Chapter II provides an overview of causal frameworks and discusses implications of randomization problems with regard to important design assumptions. A brief history of randomizations, its advantages, and the potential randomization threats are discussed. In addition, a description of statistical strategies for correcting the effects of randomization disruption is provided. Chapter III describes the procedures and use of the longitudinal data from the national Early Head Start Research and Evaluation Project as the basis for investigating the different conditions of randomization disruption and the subsequent effects on study findings. Chapter III also provides an explanation of the key study variables and addresses the statistical and analytical plans to answer the research questions.

CHAPTER II

REVIEW OF LITERATURE

Overview

Chapter II provides a review of the conceptual and research literature associated with the current methodological study. It begins with a brief history of the theoretical frameworks in which this study was housed. The study links evaluation theory within a causal framework in the context of the field of education, its relationship with experimental design, and the linked assumption of strongly ignorable intervention assignment. Then, I describe the role of randomization in experimentation, its origins, purpose, and advantages. Next, I discuss the supporting literature for the main premise of the study: threats to randomization and their implications for causal evaluation research in education. This discussion provides a brief overview of different types and causes of randomization threats and their consequences and contains a more thorough focus on the threats under investigation in the current study.

The three types of randomization threats I investigated in this dissertation study, which are discussed in greater detail in this chapter, include biased allocation, intervention noncompliance, and differential attrition. After this section, a review of methods that have been used to identify/recognize randomization threats is provided along with a summary of corrective statistical methods used in the literature to ameliorate the problems associated with failed design assumptions: analysis of covariance, propensity score analysis, complier average causal effect analysis, and multiple imputation.

Finally, the chapter concludes with a description of the Early Head Start Research and Evaluation Study (EHSRE) that served as the secondary data source employed in the current study. I provide an overview of the goals, structure, and general findings of the EHSRE study.

Causality and the Role of Randomization

Causal Frameworks, Counterfactuals, and the Assumption of Strong Ignorability

Guo and Fraser (2010) described program evaluation as the study of cause and effects (p. 21). They explained that causality in an evaluation context is the net gain or loss in the outcome of the intervention group that can be attributed to the intervention. The aim of clearly identifying causes and effects is central to intervention studies (Rubin, 1972; Ward, 2009). Causal and counterfactual frameworks provided the conceptual basis for this research study (Guo & Fraser, 2010; Heckman, 2005; Holland, 1986; Rubin, 1972). Heckman (2005) described causation in relation to the stability of the effect of intervention. If, in holding all factors constant except one (i.e., the intervention), an effect or change occurs, it is considered a causal effect (Heckman, 2005, p. 1). Sekhon (2008) described causal effects as the difference between two potential outcomes; however, only one of the two potential outcomes is observed. This is a reference to the outcomes resulting from the intervention and the outcomes that would have resulted in the absence of the intervention (called the counterfactual).

The origin of the counterfactual framework is attributed to Neyman and Rubin (Sekhon, 2007). Guo and Fraser (2010) described the counterfactual framework as the way in which causality is investigated. The counterfactual is the potential outcome if the participants had not received the intervention (Guo & Fraser, 2010, p. 24). Control or comparison groups are often referred to as counterfactual groups. The counterfactual framework is helpful in considering the purpose and benefits of randomization. The counterfactual group is meant to represent what *would have happened* to the *same* intervention participants if they had *not* participated in the intervention. It stands to reason that we would want the counterfactual group to be as close to identical to the intervention group as possible. This forms the basis of the rationale for establishing baseline group equivalence prior to the intervention. Holland (1986) suggested that "the language and framework of experiments" is the model for causal inference (p. 946).

Related to this, the concept of strong ignorability is an essential premise in estimating causal effects in observational studies (Rosenbaum & Rubin, 1983; Steiner et al, 2010). Strong ignorability is a critical treatment assignment assumption that Rosenbaum and Rubin (1983) described as met when assignment to the intervention group or the control group is independent of the potential outcomes, holding all covariates constant. When the assumption of strong ignorability in treatment assignment is met, the outcome results are unconfounded or have little or no bias (error; Pearl, 2010; Shadish & Steiner, 2010). If the treatment assignment is balanced on all covariates, then it is considered strongly ignorable (Emura, Wang, & Katsuyama, 2008; Pearl, 2010; Shadish & Steiner, 2010). Violations of strong ignorability lead to biased estimates of average treatment effects. Testing the assumption of strong ignorability is described as a statistical issue of considerable importance but little literature addresses how to assess the assumption (Emura et al., 2008). No "supreme test" of the assumption exists according to Will Shadish (Personal communication, August 10, 2011). When the assumption of strong ignorability is violated in a randomized design, it means that selection bias is present and the benefits of the design might diminish dramatically. When group assignment is not strongly ignorable, the intervention and control groups might not match on important characteristics; thus the control group no longer serves as a good counterfactual.

Origin of Randomization

Experimentation originated in the physical sciences; its success was due to the fact that the study of physical matter was amenable to the context of a controlled laboratory setting (Gall et al., 2003). In Fisher's (1925) first work clarifying his position on experimental design--*Statistical Methods for Research Workers* (with many reprint editions in subsequent years), he claimed the apotheosis of randomization, although he only discussed experimental design briefly in the book. His experiments and connections with researchers in agriculture and biology shaped his view on the benefits of randomization (Fisher, 1925, 1935; Hall, 2002, 2007; Yates, 1951). Hall's (2002) dissertation, an exposition on randomization and its history, focused primarily and comprehensively on the role of Fisher. She attributed the change from systematic design to the randomized design of experiments that occurred in the first half of the 20th century to Fisher. Mention of earlier experimental design is also present in Hall's work. She cited the research of Peirce and his student Jastrow, who are known to be the first to randomize in experiments in the late 1800s. However, the practice presumably halted until Fisher's work (Hall, 2007).

Fisher's (1935) advocacy for randomization grew from his research with small samples and understanding the role randomness plays with regard to sampling. He reasoned that randomization eliminated experimental bias and enabled tests of significance to be performed (Hall, 2002). Twenty-five years after *Statistical Methods for Research Workers* was published, Yates (1951) described the vast influence Fisher's work had on scientific research, terming it a "complete revolution" (p. 19). However, according to Hall's (2002) research, it was not initially well received. Fisher's special contribution to experimental design was the development of analysis of variance and his strong view of the necessity of randomization, which would help ensure that the "estimates of error and tests of significance should be fully valid" (Yates, 1951, p. 26). Yates believed the use of experimental design was increasingly appropriate given its benefits in improving the accuracy and certainty of experimental results (p. 33). Later work by Fisher was characterized by more specificity around how to design and conduct randomization and replication; for example, he published a historic work called *The Design of Experiments* in 1935.

With a different take on the history of social experiments (i.e., those conducted outside a laboratory), Dehue (2001) provided an historical reconstruction of an earlier and more gradual influence of the field of psychology on the randomized controlled design. She claimed that as early as the 1870s, psychologists (or as she called them, psychophysical researchers) were deliberately forming intervention groups with control comparisons (p. 289). In terms of the field of education, experimentation came into play in some of the early 1900 studies of school children (Dehue, 2001).

The work of Campbell and Stanley in the 1960s and beyond provided deeper examples of the types of experimental and quasi-experimental designs, their relative contributions to internal validity, and implications of experimentation for policy (Campbell, 1969; Campbell & Stanley, 1963). A 1994 article from Cook and Shadish reviewed developments in social experiments across multiple disciplines including psychology and education over the preceding 15 year period. They claimed a substantial change in the "dominant theory" of social experimentation during this time was characterized by an increase in the use of causation theories and more specific definitions and priorities regarding internal and external validity. Some advances in quasi-experimentation and concerns about generalized causal inferences were also discussed (Cook & Shadish, 1994). They also highlighted key aspects of designing and maintaining a randomized experiment, which were of central interest to the current study and are addressed below. Also of relevance, Cook and Shadish stated that much of the discussion about randomized experiments during this time was related to implementing them better and more often (p. 557). This has continued to be the state of affairs in recent years.

Purpose of Randomization

Experimental designs, i.e., research studies employing randomization, are the most powerful quantitative research method of establishing a causal relationship between two or more variables (Fisher, 1925, 1935; Gall et al., 2003; Hall, 2007; National Research Council, 2002; Steiner et al., 2010). In education, the purpose of using random assignment is primarily to help determine whether educational practices, interventions, and strategies have a true effect on recipient outcomes (recipients being, for example, students, teachers, administrators, or parents taking part in the intervention). In other words, using random assignment helps researchers answer research questions that seek whether and to what extent an intervention actually causes positive outcomes for recipients of the intervention while controlling for the effects of other potential causes. In the absence of random assignment, confounding or extraneous variables might lead to incorrectly attributing change to the intervention when it was the extraneous variables that led to the difference. Random assignment controls for confounds such as personal or demographic characteristics that might lead to differences at posttest.

Quasi-experimental or correlational studies (i.e., those not using random assignment) might suggest causal relationships; however, these could be misleading or altogether incorrect. Gall et al. (2003) described an example in education in which correlational findings suggested a causal relationship that had implications for modifications to teacher practice. However, an experimental study found the relationship was not corroborated.

Clearly though, the purpose of random assignment is *not* to answer all types of research questions in education. Random assignment is not meant to answer questions such as “How do students develop a sense of school belonging?”, which is better addressed by qualitative methods, or “What are the mechanisms or factors that mediate the relationship between organizational structure and teacher perceptions of support?”, which is more likely investigated by other quantitative designs using statistical techniques such as structural equation modeling. Random assignment is not meant to be used to answer all research questions--only those that are focused on determining whether treatments or interventions work.

Advantages/Benefits of Randomization

The chief advantages of employing random assignment include eliminating selection bias attributable to pre-existing group differences, controlling for confounds or extraneous influences on the outcome under study, and upholding internal and external validity of study findings. Again, random assignment controls for diverse selection processes that might influence whether the comparisons of intervention and control groups are valid (Ong-Dean et al., 2011). Random assignment offers a unique power in an intervention study in answering the question, "Does it make a difference?" (Gueron, 2001, p. 15). The randomness of the process is what helps produce a control group that can be used as a convincing and unbiased estimate of the counterfactual, i.e., what would have happened to the intervention group had they not received the intervention (Gueron, 2001, p. 18). "When perfectly implemented, random assignment generates unbiased causal estimates because at pretest the intervention and control groups are equivalent on expectation over all possible covariates" (Steiner et al., 2010, p. 250). A related advantage of random assignment is that it enables researchers to "make inferences without modeling assumptions" (i.e., assumptions such as equivalent selection; Papineau, 1994, p. 448), which means we can claim a causal relationship. Random assignment has the ability to provide control not only for confounds we know about but it is especially helpful with respect to "nuisance" variables we do not know about (Papineau, 1994). In random assignment, all such influences are probabilistically independent of the intervention.

Randomization Threats and Disruption

Threats-in-General

Failures of random intervention assignment account for many of the problems in large-scale experiments (King, Neilson, Coberley, Pope, & Wells, 2011). Threats to randomization include a whole host of problems from the beginning to the end of a study. If randomization is threatened and ultimately disrupted, then the aforementioned advantages are likely to be lost. Gueron (2001) indicated it is an "all-or-nothing process" and that the design cannot be "a little bit" randomized; "once the process is undercut, the study cannot recover" (p. 26). In this current study, I aimed to test whether at least partial recovery from randomization disruption was possible. While experiments are a powerful research design, they are not perfect (Gall et al., 2003). Popper (1968) further explained that no single experiment proves cause and effect and that experiments must be replicated. Random assignment might be threatened during one or more general study phases including (a) design of random assignment strategies, (b) allocation of participants to assigned group, (c) intervention and data collection, (d) data and randomization checking, and (e) data analysis. Downs et al. (2010) described practical problems in implementing randomization from the basis of their collective experience illustrated in clinical case examples. They indicated that at least three types of problems could occur in randomization: the first pertained to judgment errors in the choice of randomization method; the second type of errors occurred during the actual implementation of the chosen method; and the third type related to human errors that occurred during the trial, specifically by those managing the randomization process (Downs et al., 2010).

Selection of an inappropriate randomization type for an experimental study is a problem that might occur in the study planning phase. The type of intervention, sample, data, and intended levels of analysis are all important considerations in the selection of a design and randomization process best suited to realize the benefits. Corroborating this, King et al. (2011) stated that random assignment is the “defining feature of modern experimental design” (p. S-11); yet errors in design, implementation, and analysis often result in the failed realization of its benefits. They discussed problems including control of variability, levels of randomization, size of intervention arms (groups), power to detect causal effects, as well as many other problems that commonly lead to post-intervention bias. Unexpected problems can arise from myriad sources such as the research participants, school officials, teachers, politicians, or others interested in affecting the assignment of people to intervention and control groups or affecting the subsequent results. It is a serious problem when researchers do not anticipate and prevent these or other issues in the design and do not later respond to them by choosing statistical methods to correct such problems (King et al., 2011).

While it is not known how frequently major errors in randomization occur, it is considered fairly rare because it is not often mentioned in the research literature. Downs et al. (2010) stated this is false and that, based on their experience, errors in the intervention allocation process are “surprisingly common” (p. 236). Dunford (1990) also indicated that many issues affecting randomization, such as ethical and legal issues, liability risks, known manipulation of allocation outcomes, and hidden threats to randomization “regularly surface” during the implementation and operation of experiments; if not addressed, they can “wreck otherwise outstanding research designs”

(p. 125). He also claimed that practical guidance for the implementation of randomization is lacking (Dunford, 1990). Even well-designed experiments could face complications such as noncompliance and missing data (Jin & Rubin, 2009). Baker and Kramer (2008) referred to such complications as real world problems, citing loss-to-follow up (attrition), missing outcomes, noncompliance, and nonrandom selection among the issues. Other potential problems similar to noncompliance include treatment diffusion in which control participants might improve on outcomes from indirect exposure to the intervention or treatment contamination (when control participants receive aspects of the intervention when they should not). Post-randomized differential attrition might “corrupt group balance” even when groups are randomized (Guo & Fraser, 2010, p. 280). Phrased another way, when intervention or control participants differentially drop out, it might cause selection bias.

Arceneaux, Alan, and Green (2004) described another type of randomization problem that could be experienced in the analytic phase of the study. By using incorrect randomization check procedures (e.g., statistical comparison of intervention and treatment groups on baseline variables), randomization failure would remain undiscovered. For example, they showed that invalid conclusions could result from conducting randomization checks at the incorrect level of analysis, e.g., checking at the individual level when randomization is at the group level, as in cluster randomized studies.

In summary, many ways exist in which randomization and its associated benefits might be disrupted. The context and challenges of research in education lead to an increased likelihood of certain types of randomization threats (Cook, 2003; Ong-Dean et al., 2004). The randomization threats selected for examination in the current study included biased allocation procedures, intervention noncompliance, and differential attrition. Each of these is described in more detail in the paragraphs below.

Biased Allocation

Randomization processes make use of random allocation or distribution of participants to the intervention and control groups. Allocation processes (i.e., how participants are assigned to a group) might not be implemented as intended. The allocation process might thus be biased in a manner that could ruin the randomization by creating an imbalance in the covariates across the intervention and control groups (i.e., introducing selection bias; Berger & Weinstein, 2004). For example, the investigators might not have provided adequate instructions to those who recruited and enrolled participants into the study. Other times, it might be that those responsible for conducting the allocation into intervention and control groups did so incorrectly even with good instructions but poor monitoring. Downs et al. (2010) described examples when those responsible for carrying out the randomization made mistakes such that the actual implementation differed substantially from the planned method. Again, such mistakes could have derived from a lack of understanding or clarity around the allocation procedures or could be related to technical problems in programming randomization algorithms (Down et al., 2010). In some cases, intentional deviations from the planned procedures might occur. Gueron (2001) explained that program staff or teachers generally tend to dislike random assignment (p. 26). She described how program staff prefer to allocate based on their own methods: allocations based on their personal values --first come/first served, serving the most in-need, or serving the most motivated or those who most desire the service. Gueron stated further, “It is critical to design the actual random assignment process so that it cannot be gamed by intake staff” (p. 28). Berger and Weinstein (2004) described another example of allocation bias in which patterns of random assignment could be predicted based on the previous assignments; this advanced knowledge could influence strategic rather than random placement into groups.

Intervention Noncompliance

Another common issue in randomization disruption is when study participants do not take part in their respective intervention assignments as intended. Intervention noncompliance might refer to situations in which participants in the intervention group did not receive the intervention at all or in the intended dosage, intensity, or frequency (Chen et al., 2009). Many educational situations exist in which participants assigned to the intervention did not actually receive the intervention as intended. For instance, children enrolled in a classroom-based intervention might have poor attendance. Parents involved in a literacy intervention might not engage with program staff in meaningful ways, thereby affecting their children’s reading outcomes. Teachers might not attend all sessions of a professional development workshop under study. Noncompliance in a random assignment study could lead to biased estimates of the intervention effects (Bijwaard & Ridder, 2005; Esterling, Neblo, & Lazer, 2011). Esterling et al. (2011) went so far to say that noncompliance could “destroy” randomization. Indeed, noncompliance might severely limit the internal validity of the experiment (Atchade & Wantcheckon, 2005). In many cases, noncompliance might be accompanied by missing data (i.e., nonresponse; Chen et al., 2009).

Differential Attrition

Differential attrition (also called mortality) is characterized by a systematic difference in the participants who drop out of the study compared to those who stay. Foster and Bickman (1996) called attrition “nonresponse in a longitudinal context” (p. 695). Participants might drop out for a variety of reasons: difficulties in complying with the research protocol, actual mortality or illness, or mobility/moving away from the study area. Some evidence exists that attrition might vary by ethnicity; however, since it is contextually dependent, it might differ study to study (Sangi-Haghpeykar, Meddaugh, Liu, & Grino, 2009). Greater attrition among some samples in education might also exist. Kubitskey et al. (2012) suggested that attrition among teachers is a new challenge for researchers given that professional development is increasingly a topic of randomized study. Another specific type of attrition is based on the participants’ level of satisfaction with or preference for their group assignment. McKnight, McKnight, Sidani, and Figueredo (2007) discussed that this division between the intervention (satisfied/not satisfied) and control (satisfied/not satisfied) groups could drive some dropout among the “disappointed” individuals (p. 30).

As part of a trial attrition study group, Hewitt, Kumaravel, Dumville, and Torgerson (2010) found that nearly 25% of randomized controlled trials had more than 10% of the study outcome data missing. They reported that balance in the baseline characteristics across groups might be ruined. Likewise, attrition can have negative effects on the internal and external validity of a study (Flick, 1988; Foster & Bickman, 1996; Marczyk, DeMatteo, & Festinger, 2005; McKnight et al., 2007; Shadish, Cook, & Campbell, 2002). Marcellus (2004) claimed attrition is underreported and understudied despite its potential to bias study findings. Hewitt et al. (2010) reported that “although attrition is common, it is unclear when it becomes a serious threat to trial validity” (p. 1265). Following this recommendation, this present study intends to add to what is known about *when* attrition threatens the validity of findings.

The loss of participants (i.e., decrease in sample size) might also lead to a loss of both statistical power and an ability to detect group differences if they exist. Further data analytic problems could result from loss of data. For instance, the intervention groups might no longer have equal sample sizes; this could contribute to violations of the assumptions of normality (distribution of error) and homogeneity of variance (McKnight et al., 2007). Inequality in analyzable cell sizes might be a large part of why attrition causes so many validity problems. Marczyk et al. (2005) suggested that if the participants who dropped out were substantially different from those who stayed, it might limit the study’s generalizability. Not only would the findings be non-generalizable in the sense of the immediate study but McKnight et al. (2007) explained that attrition might negatively influence the synthesis of results across studies and hinder the knowledge base in the field and associated theory development (p. 35). Fitzmaurice (2003) indicated that the attrition is nonignorable when the likelihood of dropping out is fundamentally related to the outcome. In this case, biased effect estimates are likely be found (Hewitt et al., 2010).

Differential attrition is essentially a missing data problem (McKnight et al., 2007). When the missingness is nonignorable, meaning that it is differential, nonrandom, and causes imbalance in the intervention and control group, then strategies must be used to address the missingness (Foster et al., 2004). Unfortunately, a well randomized study cannot prevent dropout; however, a well-designed and monitored study might. Purposeful follow-up with participants to maintain buy-in and rapport, individualization of retention strategies, and provision of substantial incentives could help stave off attrition and missing data issues.

Implications of Randomization Disruption

To reiterate, random assignment disruption in its various forms might diminish a study's internal and external validity. Campbell and Stanley (1963) provided a thorough review of the factors affecting the internal and external validity of experiments. They described eight different classes of extraneous variables that might result in confounds with the effects of the intervention: validity threats due to history, maturation, testing, instrumentation, statistical regression, differential selection bias, experimental mortality (or differential attrition), and selection maturation interaction. Gall et al. (2003), for instance, effectively described these and a few more from Cook and Campbell (1979): experimental intervention diffusion, compensatory rivalry by the control group, compensatory equalization of interventions, and resentful demoralization.

Threats to randomization equate to threats to validity and are subsequently associated with biased intervention effect estimates. Thus, the implications of failed randomization are clear: researchers might lose the ability to rule out any other potential causes of change on the posttest beyond the intervention effects.

Identification of Randomization Problems

Methods of identifying random assignment problems include proactive and diligent monitoring of random assignment processes and the performance of randomization checks that statistically examine group differences after baseline data collection. Problems could be identified by researchers who closely monitor the random assignment processes. This might involve masking the allocation sequence, conducting site visits for quality control, checking the dates of randomization to see if reasonable numbers of individuals are randomized within a given time period, or whether participants are randomized on unexpected days or patterns.

Randomization checks are also known as balance tests. Randomization checks include standard statistical tests to examine differences between the intervention and control groups on baseline variables, which were preferably collected prior to randomization, so the randomization process and the knowledge of assignment do not influence the measurement (Cook & Campbell, 1979; Marczyk et al., 2005). These procedures generally include comparing intervention and control groups on demographic variables such as gender, ethnicity, level of education, and other participant characteristics. In addition, baseline outcome measures might be compared. If differences are found on variables, Marczyk et al. (2005) suggested that researchers should examine whether those variables are correlated with the outcome variables and, if so, they should be controlled for in the final analyses.

Mutz and Pemantle (2011) stated that randomization checks serve a useful purpose when assessing threats to the execution of the randomization procedures. They claimed that these checks are the most important to conduct when there is some reason to believe the randomization did not work properly (p. 3). Another form of randomization check to perform when attrition is evident is to (a) compare the differences between the remaining intervention and control participants and also (b) compare the dropouts with those who remained in the study on those baseline measures to identify whether attrition was, in fact, differential.

Corrective Techniques for Selection, Noncompliance, and Attrition Biases

In this section, I briefly describe methods researchers might use to correct bias across a variety of designs under a number of conditions. The goal of these techniques is uniform around increasing the validity of study results and interpretations but the methods might reach the goal in different ways. Later in the section, I more specifically discuss the use of analysis of covariance, propensity score analysis (PSA) for reducing selection bias, compliance average causal effect (CACE) for addressing noncompliance, and multiple imputation (MI) for tackling the nonrandom missing data associated with differential attrition as these techniques were considered for use in this study. Note that within this investigation, it was not my intention to disregard the fact that avoidance and prevention of systematic errors through careful research design and monitoring was much preferred to the use of post hoc statistical corrections. Rather, the idea was to explore ways to salvage flawed studies.

The most common strategy when a variable has failed a balance test (i.e., has shown a statistically significant difference at baseline between intervention and control groups which might lead to bias) is to include that variable as a covariate in the statistical analyses (Mutz & Pemantle, 2011). However, Mutz and Pemantle (2011) asserted that the literature "reveals considerable divergence of opinion on the use of correcting for imbalance in experimental designs" (p. 9). This is due in part to the fact, even in randomization, that some imbalances are likely due to chance and some corrective procedures might overcorrect the problem. In addition, groups might differ on baseline measures of outcomes. Other strategies to ameliorate problems associated with imbalance are post-stratification (e.g., weighting the intervention and control samples prior to outcome analyses to force them to coincide/equalize on some imbalanced variable or by conducting analyses by subpopulation) or re-randomization (i.e., this can be done if balance checks are performed before the intervention begins and before participants are notified about inclusion in the intervention; Hansen & Bowers, 2008; Mutz & Pemantle, 2011).

To correct selection bias in quasi-experimental and other designs, often sample matching techniques are employed. Treatment and comparison groups might be matched prior to outcome analyses on one or more characteristics such as gender, age, or ethnicity. Propensity score analysis (PSA) is a technique used to reduce selection bias in primarily quasi-experimental studies but can be applied in a variety of designs (Clark, 2008). The propensity score is the probability of receiving treatment that is predicted by a set of covariates preferably linked to both selection and outcome. The propensity score can also be used as a composite covariate to match samples, hence propensity score matching. Propensity score analysis is discussed more thoroughly below.

To correct for intervention noncompliance, one consideration is whether the noncompliance co-occurs with nonresponse (missing data). In the presence of missing data, different strategies or combinations of strategies might be used to attempt correction (Chen et al., 2009; Dunn et al., 2005; Jin & Rubin, 2009, Little & Yau, 1998; Taylor & Zhou, 2009). In many cases, researchers might attempt to adjust for the compliance processes in analyses to be able to make a correct counterfactual comparison between the intervention and control groups (Esterling et al., 2011). This means that the variables associated with participants' compliance level might be identified and used as controls in the statistical analysis. Observing the compliance process might involve an examination of the measured covariates, e.g., demographic characteristics for those who complied versus those who did not, to see if there were differences between the two groups. Researchers might also try to model or predict participants' latent compliance type from other behavioral measures used in the study. For instance, researchers might use latent class analysis to analyze patterns of factors associated with compliance type, e.g., participation in program activities and level of involvement, and thereby categorize participants by their type of compliance (i.e., group membership; Hagenaars & McCutcheon, 2002). Commonly though, the compliance process is driven by unmeasured variables, i.e., it is unknown what led to compliance variability. For example, consider a situation in which measured pretest characteristics might not be related to dropout such as perceived incompetence of the intervention provider. A group of participants might leave because they are not benefitting but might not disclose their dissatisfaction.

The instrumental variables method (IV) is widely used to address the problem of noncompliance with the intervention when the compliance process is unobservable (Angrist, Imbens, & Rubin, 1996; Angrist & Krueger, 2001). Instrumental variables is a method commonly used to estimate causal effects when it is not possible to conduct a randomized experiment. In the instrumental variables method, a variable (called an instrument) that is correlated with the intervention but does not directly lead to change in the outcome is used to control for confounding (Greenland, 2000).

Usually researchers tend to conduct intent-to-treat (ITT) analysis in which noncompliance is essentially ignored and groups are analyzed as randomized (Hewitt, Torgerson, & Miles, 2006). Intent-to-treat analyses retain the original assignments of cases regardless of whether they received the intervention or not. The rationale is that by not doing so, worse bias would be introduced to the findings by compromising the randomization. In ITT, noncompliance is philosophically seen as just part of the response to the intervention (Atchade & Wantcheckon, 2005; Baker & Kramer, 2007); in this light, some researchers believe a person's compliance is not meaningfully separable from the effectiveness of the intervention. However, Atchade and Wantcheckon (2005) indicated that ITT estimates are reported to be highly biased compared with the real intervention effects, especially in situations when the noncompliance patterns change over time (or are really severe). However, standard recommended practice, in at least the medical field, is to conduct ITT analyses and then also provide results from per-protocol (compliers only) or on-treatment (treated only) analysis to estimate treatment effects (Hewitt et al., 2006). The problem with per-protocol and on-treatment analyses is that selection bias is introduced by abandoning the randomization. Another approach to handling noncompliance is called complier average causal effect analyses and is described below.

Statistical techniques for handling missingness in differential attrition include a wide array of strategies from simple to complex. Choice of technique is often guided by the type of analyses required to address the research questions and whether the missingness is random or systematic. Strategies briefly include data deletion (the common listwise or pairwise deletion), data augmentation (a model-based procedure such as maximum likelihood), and single and multiple imputation techniques (McKnight et al., 2007). Multiple imputation is described in more detail below as it is a common technique for addressing missingness due to attrition.

Analysis of Covariance

Another common method of handling biases or confounding in a statistical comparison of intervention and control groups is to include additional baseline explanatory variables (also called covariates) in the analysis (Clason & Mundfrom, 2012; Little, An, Johanns, & Giordani, 2000; Taylor & Innocenti, 1993). When comparing groups, this adjustment method is called analysis of covariance (ANCOVA). Analysis of covariance is considered an extension of ANOVA that adjusts the estimates of intervention and control mean differences (Raykov, 2010). Analysis of covariance is also described as a combination of ANOVA and regression (Cochran, 1957; Taylor & Innocenti, 1993).

Researchers cited numerous benefits of using ANCOVA: enhanced potential for increasing the power to detect findings, the ability to detect and estimate interactions, and improved handling of measurement error (Egger, Coleman, Ward, Reading, & Williams, 1985; Forsythe, 1987; Little et al., 2000; Taylor & Innocenti, 1993). Taylor and Innocenti (1993) stated that ANCOVA could be helpful in the context of a randomized design even when careful attention was paid to allocation procedures. In experiments, the statistical method reduced unexplained (within-group) outcome variance and thus increased the power of the intervention effects (Van Breukelen, 2010). The use of ANCOVA has been known to increase the precision of intervention estimates in randomized designs (Cochran, 1957; Liu, 2012). Liu (2012) indicated that the addition of a covariate could account for a portion of the variance in the residual error, thereby reducing the standard errors of the effect estimate. Egger et al (1985) suggested that in clinical trials, ANCOVA is most often used when baseline differences between intervention and control groups are significant but made the recommendation that it be used more routinely to adjust for confounding. However, they cautioned that the use of ANCOVA with very small samples could further bias effect estimates.

Propensity Score Analysis

Propensity score analysis is increasingly used within many design types to reduce selection bias. It is not considered a panacea or permanent replacement for an experimental randomized design but is an advantageous method when randomization is impractical or not possible. Careful attention to the limitations of the method is strongly warranted and omnipresent in the literature. For instance, Mitra and Heitjan (2007) stated, “No adjustment method, even propensity scores, can completely eliminate potential bias from imbalance on covariates that do not appear in the data set” (p. 1399).

Created by Rosenbaum and Rubin in 1983, propensity scores are the conditional probabilities of selection into the intervention group. This means that they represent how likely it is that a participant would be selected either into the intervention or the control group. Propensity scores are constructed using variables that explain the selection process so that selection bias might be controlled for in analyses. The scores are predicted (often using logistic regression) using an array of covariates correlated with both selection and outcome (Clark, 2008). Their values range from 0 to 1; scores above .5 predict being in one group (intervention) and scores below predict being in the other. The scores can be used in a variety of procedures including matching, covariate adjustment, and weighting (Guo & Fraser, 2009). In propensity score matching, the propensity score (as if it is a composite of many covariates) is used to match participants from the control and intervention groups to create group balance prior to outcome analyses. It is intended to mimic randomization by controlling selection bias.

In propensity score analysis, it is absolutely critical to know and measure the selection process, i.e., how subjects are selected into study groups (Shadish & Steiner, 2010). In Shadish and Steiner's (2010) guide to propensity score analysis, they described the iterative process of choosing the best combination of covariates to maximize bias reduction. Lack of access to unmeasured covariates contributes to the biggest challenge in using propensity score analysis (i.e., cannot adjust with variables that were not measured). Steiner et al. (2010) investigated how different sets of covariates performed in terms of minimizing selection bias in observational studies. They theorized that "…the most important covariates for supporting strong ignorability were those closely related to both the real selection process and study outcomes" (Steiner et al., 2010, p. 250). They found this was indeed the case--unrelated covariates did not substantively reduce selection bias and the actual statistical adjustment method did not particularly matter.

The goal of propensity score analysis is to imitate the baseline covariate balance between intervention and control groups achieved through randomization. Such balance is accomplished when the distributions of the baseline covariates are nearly identical for intervention and control, and likewise for the two distributions of propensity scores (Shadish & Steiner, 2010, p. 21). Propensity score analysis does this by controlling for the selection process.

Complier Average Causal Effect Estimation

Complier average causal effect (CACE; Angrist et al., 1996) estimation, also called the local average treatment effect, is the measured causal effect of the intervention among compliers randomized to the intervention group. It is used to improve causal intervention effect estimates by reporting the intervention effects for those who both received the intervention and were compliant. The influence of those in the intervention group who did *not* receive the intervention is removed from the estimate. When noncompliers are included in the estimates, they have what is called a downward bias, meaning that the mean effects are attenuated. Complier average causal effect estimation (CACE) corrects for this bias. The advantage of using CACE, compared with per-protocol (assessing complier only) and on-treatment (assessing treated only) analyses, is that CACE retains the original random assignment of the participants (Hewitt et al., 2006, p. 347). Hewitt et al. (2006) still recommended using this method in combination with ITT analyses but preferred CACE to the bias introduced by other analyses that go against the randomization. Dunn et al. (2005) emphasized Angrist et al.'s (1996) assertion that compliers who were randomly assigned to the treatment were the only group whose average treatment effects could be viewed without bias (p. 376). Essentially in CACE, the outcome results for compliers who were randomized to the intervention group are compared to the same proportion of participants estimated to be compliers in the control group (i.e., given randomization the assumption is that the distribution of compliers is the same across groups). Complier average causal effect estimation results are typically also compared with the ITT results to assess the relative improvement in the estimation of treatment effects. The analysis section in Chapter III provides a description of the step-by-step approach for using CACE. Another way to boost the effectiveness of this method, in providing a more valid (less biased) estimate of treatment effects, is to add a covariate to the analysis that explains compliance behavior (Hewitt et al., 2006). The goal with CACE is to improve estimates of average treatment effects affected by noncompliance.

Multiple Imputation

Schafer (1999) described imputation as the practice of "filling in" missing data with plausible values (p. 3). McKnight et al. (2007) cited many champions for multiple imputation for the statistical handling of missing data. They described it as easy to implement and effective in producing sound parameter estimates including the standard errors (McKnight et al., 2007, p. 196). In addition, multiple imputation can provide information about the effect of missing data on the results beyond just knowing the proportion of data missing. Clearly, higher rates of missing information have a more biasing effect on study results (Barnes, Lindborg, & Seaman, 2006; McKnight et al., 2007). Multiple imputation was first proposed by Rubin in 1977 and can be generalized to a variety of datasets and statistical problems (McKnight et al., 2007; Schafer, 1999). It is known to be statistically valid when large samples are used for treating missing data (McKnight et al., 2007; Rubin, 1977); although Barnes et al. (2006) tested it successfully with some small clinical samples. Sassler and McNally (2003) suggested that the patterns of attrition should be examined carefully to be able to effectively identify variables to include in the imputation.

Barnes et al. (2006) summarized multiple imputation succinctly by indicating three steps: imputing missing values M times, analyzing the M imputations, and combining the M analyses. Different imputation routines could be used (McKnight et al., 2007; Sassler & McNallyy, 2003) such as random normal values and hot deck values. McKnight et al. (2007) provided concrete guidance on the step by step process including the additional step of calculating the missing information estimate; these steps are presented in the data analysis section of Chapter III. The imputations are computed using an iterative process from the observed data deemed to be the most "information rich" including important covariates and other variables that can provide additional information about the variability of the data and potential explanations of the nonresponse. Rubin (1977) referred to the likelihood of being a nonrespondent as a "probabilistic function" of these important predictive background variables (p. 540). In other words, multiple imputation is used to create a small number of independent draws of the outcome variables from a predictive distribution (Schafer, 1999). Multiple imputation is considered to be appropriate to use in many different missing data situations. Little and Rubin (1987) described the classification of missing data that includes three different mechanisms (i.e., not explanations) by which data might be missing. When outcome data are considered to be missing-completely-at-random, the reason for the missingness is due to random processes and not related to the outcome, i.e., there are no systematic patterns in how the data are missing. In the case of missing-at-random, the missingness is potentially a function of some independent variables; however, correct inferences might be made if explanatory variables are used as controls. Missing-completely-at-random and missing-at-random missingness is considered ignorable. Missing-not-at-random (MNAR) is considered to be nonignorable with the probability of missingness dependent upon the outcomes in a non-predictable way (Little & Rubin, 1987; McKnight et al., 2007). Multiple imputation is used most often in missing-at-random situations but has been found to be flexible enough to be effective in missing-not-at-random conditions (McKnight et al., 2007).

Early Head Start Research and Evaluation Project

An overview of the Early Head Start Research and Evaluation Project (EHSRE) is found on the Administration for Children and Families (U.S. Department of Education, 2002a, 2002b) website. The initiative, funded through the Office of Planning, Research, and Evaluation, is described as a rigorous, large-scale, random-assignment evaluation of Early Head Start (EHS). It was designed to carry out the recommendation of the Advisory Committee on Services for Families with Infants and Toddlers for a strong research and evaluation component to support continuous improvement within the EHS program and to meet the 1994 reauthorization requirement for a national evaluation of the new infant-toddler program.

The EHSRE study waves were intended to examine whether the programs' child development, parenting, and family development services were effective in promoting positive parent and child outcomes. The study was designed and carried out by the EHS Research Consortium. Program services were tailored to the needs of low-income pregnant women and families with infants and toddlers. The families were diverse (U.S. Department of Education, 2002a, p. 25) and program activities were particularly effective in improving child development and parenting outcomes of African American participants (U.S. Department of Education, 2002a, p. 8). Program approaches included center-, home-, and combination-based services. While the EHS programs across the 17 sites used different approaches for delivering their services to children and families, all approaches produced positive impacts on child and parent outcomes (U.S. Department of Education, 2002a).

Early Head Start promoted numerous positive child and family outcomes. “Overall impacts were modest, with effect sizes in the 10 to 20 percent range, although impacts were considerably larger for some subgroups, with some effect sizes in the 20 to 50 percent range” (U.S. Department of Education, 2002a, p. 3). Statistically significant gains on standardized assessment of cognitive development were found at age 2 and sustained at age 3. In addition, the program intervention group scored significantly higher than did the children in the control group. Likewise, significant positive effects were also found for language development and for other social-emotional development indicators such as aggression. Administration for Children and Families (U.S. Department of Education, 2002a) also found evidence that the positive effects on children when they were 3-years-old were associated with positive changes in parenting when children were 2-years-old, consistent with the programs’ theories of change (p. 5).

Positive gains were also made toward family self-sufficiency with significant positive outcomes related to participation in education and job training activities. The intervention and study also had a strong fathering component. While the program did not show a positive effect on maternal depression (there was a very high rate of depression among EHS families), mothers in the program group showed improved parenting interactions and relationships with their children (U.S. Department of Education, 2006).

The EHS sites each had a local research team that helped carry out the national data collection as well as the data collection for the local research questions. Numerous supplemental studies and associated papers, symposia, and reports were produced from these national and local data. One such study by Raikes et al. (2006) addressed the question of "Under what conditions is home visiting an effective service strategy?" (p. 3). They identified a number of predictors of home visiting involvement and the linkages with child and parent outcomes. The findings from Raikes et al. and other articles and reports helped guide the selection of variables (described in greater detail in Chapter III) for use in the current study.

Summary

It is disappointing that even the greatest design available for estimating causal effects is so commonly susceptible to disruption. The elegant benefits afforded by randomizing research participants to groups, e.g., achieving balance to ensure baseline equivalence, might be thwarted by other challenges encountered during the study. If only time travel were possible so researchers could go back to observe outcomes for the intervention participants for whom intervention could also be withheld! Then a truer understanding of the efficacy of the intervention would be made clear when the effects of treatment and the effects of no treatment are studied using the exact same subjects. Challenges such as allocation bias, intervention noncompliance, and differential attrition in different contexts pose different degrees of threats to internal and external validity by creating biases that were hoped to be avoided by using the randomized design. Certainly researchers might take steps to avoid bias by more careful monitoring of processes, partners, and participants. In the event problems do occur, it is important to have an arsenal of effective, corrective techniques to combat threats to the power of randomization.

It is clearly known that randomization disruption might create problems in a research design that leads to invalid findings. What is not precisely known is how much disruption must be present to bias findings. In other words, it is not well understood how robust randomization is to common threats and how sample size affects the magnitude of disruption.

Chapter III provides a description of the current study's design, data source, variables, procedures, and analysis intended to shed light on the degree of threat common study problems pose to randomization using a national set of early childhood educational intervention data. A selection of corrective techniques and the resources used to carry out the procedures is described.

CHAPTER III

METHODOLOGY

The design, data source, variables, procedures, and analyses for this study are described within this chapter. A detailed description of the randomization threat scenarios and associated conditions are also provided along with an explanation of how these conditions were constructed with and applied to the Early Head Start Research and Evaluation Project (EHSRE) datasets for comparative analyses.

Design

To address the research questions, this methodological study employed a comparative, multi-condition design to examine differential effects of randomization threats on randomization and on EHSRE study outcomes. I also compared the relative efficacy of two corrective techniques in restoring data findings when randomization was disrupted by imposed threat conditions. In addition, I also examined whether the corrective techniques, under some conditions, might further distort the results found under threat conditions.

Fifty-four different datasets were created, each representing 1 of 27 randomization threat conditions and created by two different methods. The 54 datasets were constructed with two methods using three types of threats representing typical study problem situations, three sample sizes representing study samples of different sizes, and three proportions of the sample exposed to the threat condition representing degree/severity/frequency of the threat. Each of these 27 conditions was tested using four different child outcome measures from the EHSRE data to see if disruption of randomization varied by outcome type. The intention of using these conditions was to increase the present study's generalizability by providing results representing a variety of potential problem types, sample sizes, and degrees of problem. Figure 1 provides a summary of the threat conditions.

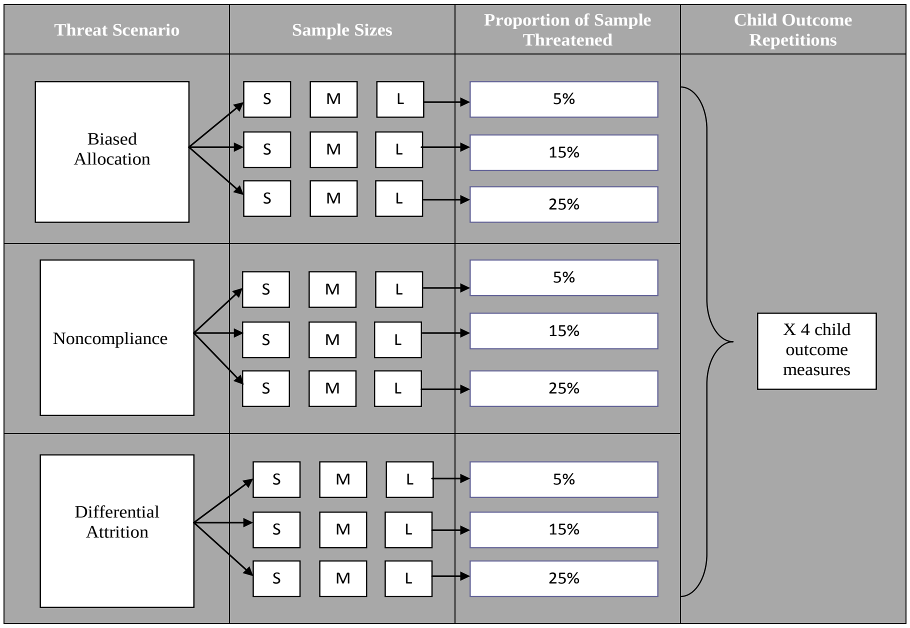


Figure 1. Randomization threat conditions.

Threat Type

The conditions were founded on three typical randomization threat types: biased allocation, intervention noncompliance, and differential attrition. The biased allocation threat involved a scenario in which the randomization procedures were disrupted and the proportion of families with higher needs was increased in the intervention group. The noncompliance threat involved a scenario in which the proportion of noncompliers in the intervention was increased. For differential attrition, the proportion of cases in the intervention sample with fewer risk factors was increased (dropout of higher risk cases). In the procedures section in the Phase II dataset construction section, I provide a descriptive scenario that includes a plausible background for the threat type and the method of dataset construction.

Proportion Exposed

The percentages of data exposed to the threat condition were intended to represent levels of increasing severity/frequency of the threat to randomization. The proportions of the sample exposed to the threat were 5%, 15%, and 25%. Foster et al. (2004) reported that it is a common goal of researchers to have no more than 20% attrition. The threat percentages applied in the current study provided values above and below that level to help learn when problems begin to arise. The proportions were also selected on guidance from Grimes and Schulz (2002) who suggested that attrition of 5% or lower was unlikely to introduce bias, while 20% was likely to be a problem. However, the effect depended on whether the attrition was random or nonrandom (Hewitt et al., 2010). Hewitt et al. (2010) suggested that a wide range of attrition levels should be studied with many different covariates to add to “rules of thumb” regarding when attrition is likely to be a problem (p. 1269).

Sample Size

To increase the generalizability of the current study’s findings, three different sample sizes were used within the conditions to understand the differential effects of the three types of threats for studies of small, medium, and large size. The large sample size conditions were based on the cases available in the original Early Head Start data. The average sample size across data collection time points for the outcome measures was approximately 1,700. To obtain the large sample datasets, 1,400 cases were randomly sampled from the original Early Head Start database with 700 in the intervention group and 700 in the control group.

The small sample size conditions were based on average small sample sizes represented in the educational evaluation literature. Slavin and Smith (2009) investigated the relationship between sample sizes and effect sizes in a systematic review in education studies. In the process of their review, they described smaller samples were comprised of fewer than 200 to 250 subjects (Slavin & Smith, 2009, p. 500). They considered very small studies to be those with sample sizes below 50. Thus, a sample size of 125 for control and 125 for intervention groups was selected based on this small sample guidance. The 250 sample size represented approximately 18% of the large sample size condition. The medium sample size condition was chosen as a midpoint between the two sizes--600 total with 300 control and 300 intervention.

Institutional Review Board Application and Data Protection

Upon acceptance of the research proposal by the University of Northern Colorado (UNC) Graduate School and prior to the commencement of the study, online application was made to obtain Institutional Review Board (IRB) review of the study’s compliance with required conduct and standards for research. The request proposed the use of a de-indentified secondary dataset not requiring new individual participant informed consent procedures. The UNC electronic IRB system was used to apply and address any subsequent IRB comments and questions. After approval, research activities for the study commenced (see Appendix A). Databases and associated documentation were stored on a secured, password-protected personal computer and measures were taken to ensure the protection of the data from unauthorized access and viewing.

Data Source

For the present study, I used secondary data originally collected during the first wave (Birth to Three) of the national Early Head Start Research and Evaluation Project (EHSRE). The original randomized study was funded by the U.S. Department of Health and Human Services, Administration for Children and Families (ACF; 2002a), Office of Planning, Research, and Evaluation. Data were collected through the Early Head Start (EHS) Research Consortium which was comprised of representatives from ACF, evaluation contractors, 15 local research teams, and 17 EHS programs. I obtained permission from the EHS Research Consortium to use the public and private EHSRE datasets for dissertation work and was provided access to the secure online repository storing the data (see Appendix B). In addition, substantive documentation files and explanatory materials were provided to help in interpretation and use of these data.

The first wave of EHSRE was conducted from 1996 to 2001 and the study files containED program, parent, and child data from both the implementation and impact evaluations. The data files were constructed and maintained by researchers and analysts from Mathematica Policy Research, Inc. (MPR, one of the evaluation contractors), who also created numerous evaluation reports describing the methods and findings for the EHSRE project.

At enrollment, as reported by ACF (2002a), the EHSRE study sample was comprised of families receiving public assistance of some kind: 77% were receiving

Medicaid, 88% were receiving Women, Infants, and Children (WIC) benefits, 50% received food stamps, and about 36% were receiving Temporary Assistance for Needy Families (TANF). Early Head Start applicants were the primary caregivers for the eligible children (99% mothers). They were on average 23-years-old and about a third were teenage parents. About a quarter of these families lived with a spouse. One-third of the families were African American, 25% were Latino, a little over a third were Caucasian, and a small percentage were other ethnicities. About 20% of the primary caregivers did not speak English as their primary language. About half of the primary caregivers lacked a high school diploma, 23% were employed, 22% were in school/training, and 55% were neither employed nor in school. About 25% of the caregivers enrolled while they were pregnant. To be eligible for the study, the research families had to be pregnant or have a child younger than 12 months old. About 50% of the children were younger than 5-months-old at enrollment, 61% were first born children, and 10% were considered low birth weight.

There were two essential prerequisites for using these Early Head Start Research and Evaluation Project (EHSRE) data for the present study: (a) the randomization for the original study was implemented properly and (b) research nonresponse did not cause randomization failure. Both of these conditions were met. MPR was responsible for oversight of the randomization process used at each of the 17 programs which involved using computer-generated random numbers to assign eligible families to a research status. They reported that the process used to randomize was implemented correctly, and that staff at the programs and local research sites followed the specified procedures (ACF, 2002b, p. D.5). Mathematica Policy Research compared the characteristics of the intervention and control group families (collected at baseline prior to randomization) to determine whether the random assignment process was implemented correctly. They found that the research groups had equivalent characteristics (ACF, 2002b, p. D.12). Given that appropriate implementation of the random assignment was a critical, *sine qua non* precursor to using these data for the current study, this was a positive finding and further justified the use of these data.

With regard to the issue of study nonresponse (i.e., whether participants complied with research activities, such as interviews and assessments), Mathematica Policy Research found, upon comparison, some differences between research respondents and non-respondents but the differences were not large and were similar for both the intervention and control groups (ACF, 2002b, p. D.31). They concluded that the resulting impact estimates were likely to be unbiased because the characteristics of the respondents in each of the research groups were similar.

Study Variables and Measures

Child Outcome Measures

Four different child outcome variables were examined in the current study to determine the comparative effects of each threat condition. The child variables examined included one from each of four developmental domains: cognitive, language, behavior, and social emotional. A variety of outcome variables were selected to help determine whether effects would be similar across domains, again supporting the generalizability of potential findings.

Child outcome variables investigated included assessment data collected when the children were 24 and 36 months of age; the variables included (a) mental developmental index (MDI) scores on the Bayley Scales of Infant Development-Second Edition (BSID-II; Bayley, 1993); (b) standard scores from the Peabody Picture Vocabulary Test-III (PPVT-III; Dunn & Dunn, 1997); (c) Child Behavior Checklist--aggression subscale scores (CBCL; Achenbach & Rescorla, 2000); and (d) the Engagement of Parent scores on a parent-child, semi-structured play interaction (Three Box coding scales; National Institute of Child Health and Human Development [NICHD] Early Child Care Research Network, 1997, 1999).

The BSID-II is a direct, individual child assessment battery comprised of five scales: cognitive, motor, language, and two parent questionnaires to assess social-emotional development and adaptive behavior. It is a norm-referenced assessment with strong evidence of reliability and validity for both research and the detection of developmental delays based on a variety of low/high income, diverse samples. The MDI was used to assess intervention versus control differences across present study conditions. The MDI is a standardized score based on national norms that take into account the infant's chronological age at the time of testing (adjusted for premature birth) to assess whether children are on track developmentally. Scores from the standardization sample were normally distributed with a mean of 100 and a standard deviation of 15 (Gauthier, Bauer, Messinger, & Closius, 1999).

The PPVT-III is a well-known and frequently used standardized measure of children's receptive vocabulary. It is also a direct child assessment and is administered using a booklet of pictures; children are instructed to point to the slide that shows the word spoken by the assessor. This test, as with the BSID-II, yields age-based standard scores with a mean of 100 and standard deviation of 15.

The CBCL, for ages 1½ to 5 years, is a parent-report questionnaire that assesses children's competencies and behavioral and emotional problems. While the full CBCL contains 118 items, only the aggression subscale was assessed for EHSRE. This subscale included 19 child behavior problem items. Scores range from 0 (no incidence of aggression) to 38 (if all behaviors are often observed).

A semi-structured, free play task was coded from videotape according to scales adapted from the NICHD (1997, 1999) study of Early Child Care's Three Box Task. Only one of the three child scales for this scheme, Engagement of Parent, was used in the current study. The scale was coded on a 7-point scale and reflected the extent to which the child showed, initiated, and/or maintained interaction with the parent. "This may be expressed by approaching or orienting toward parent, establishing eye contact with parent, positively responding to parent's initiations, positive affect directed to parent, and/or engaging parent in play" (ACF, 2002b, p. C-10). Table 1 shows the outcome variables by construct, domain, and method of data collection.

According to ACF (2002b), the child measures were originally selected for the EHSRE project based on the following guiding principles: (a) relevance to intervention goals and hypotheses, (b) appropriateness to children's age and developmental level, (c) appropriateness for the EHS population, (d) adequate psychometric properties, (e) prior use in large-scale intervention evaluations, and (f) low cost and burden. These variables were selected for the current study from the 24- and 36-month data collection age points. However, the PPVT data were not collected at 24 months so the analysis for that one variable was point-in-time only (see analysis section below).

Table 1

Child Measures and Outcomes by Domain and Collection Method

Measure	Outcome Variable	Developmental Domain	Method
BSID-III	MDI Score	Cognitive	Individual Assessment by Researcher
PPVT-III*	Standard Score	Receptive Language	Individual Assessment by Researcher
CBCL	Aggression Subscale Score	Behavioral	Parent-Report Survey
Semi-structured Play Interaction	Engagement of Parent Score	Social-Emotional	Observational Coding by Researcher

* Collected at 36 months only.

These outcome measures were selected because they represented different developmental domains. The careful selection of age-appropriate and high psychometric quality measures for the original study further supported the justification for their use in the current study.

Family and Child Baseline Characteristics

Baseline characteristics were used to gauge the equivalence of the intervention and control groups after data were subjected to each threat condition. Thus, this set of variables was used to test the degree of disruption of randomization. For consistency, I used the EHSRE's 12 key variables for establishing research group equivalence used in the original study. Family variables included age of mother at random assignment, mother's level of education, race and ethnicity, primary occupation, English language ability, living arrangements, household income as a percent of the poverty level, public assistance receipt, food stamps receipt, and maternal risk index. Child baseline variables included age at enrollment, gender, and born more than three weeks early. These variables were also used in Phase IV of the present study to assess the efficacy of the corrective procedures.

Early Head Start Attendance and Service Use

The degree of attendance, involvement, and service use in the Early Head Start (EHS) service were important variables for the current study. They were used to help create datasets subjected to threat conditions under the noncompliance scenario. I created a constructed variable of overall involvement in program activities and used it to identify/characterize families along the continuum of participation. The individual variables used to measure participation and collected with the project’s 26-month parent services interview included (a) length of participation in EHS in months, (b) frequency of participation in EHS activities, (c) time per month parent spent in EHS activities (hours), and (d) time per month the focus child spent in EHS activities (hours) including child development services. The EHSRE study created composites regarding participation based on patterns of these variables.

Procedures

The project had four primary procedural phases. Phase I operated as a pilot to serve as both a data preparation stage and a preliminary run-through of one full randomization threat condition. During Phase II, data were generated to specification for each of the remaining threat conditions. Phase III addressed Research Questions 1 and 2 by performing diagnostic and comparative sensitivity analyses to assess differences in each study condition compared to findings from the pure, randomized sample. To address Research Question 3, Phase IV involved examining the comparative performance of data analytic techniques to attempt corrections or adjustments of data to improve the accuracy of the results.

Phase I: Data Preparation and Pilot with One Threat Condition

Phase I (preparation and pilot) was used to conduct the subsequent phases of the study in the most efficient and effective way. It was anticipated that Phase I would provide an opportunity to (a) gain a deeper understanding of the existing data structure and the relationships among the study databases and variables, (b) construct the foundation datasets, (c) pilot a single condition, and (d) refine the covariate list for propensity score analysis and other adjustment methods used in Phase IV of the study.

Foundation datasets. Since there were 20 different databases with original EHSRE data, I first merged together only the variables to be used in the study to construct a new, parsimonious, easier-to-manipulate dataset. From this new original base set (OBS), I constructed the large-sample foundation dataset (labeled FL for Foundation Large) that randomly sampled 1,400 cases from the OBS (i.e., 700 intervention and 700 control). I created the medium sample foundation dataset (*N* = 600; FM for Foundation Medium) by randomly selecting cases equally from the OBS intervention and control groups. I also created the small sample-sized dataset (*N* = 250; FS for Foundation Small) by randomly selecting cases equally from the OBS. These steps formed the three foundation groups and retained the randomization and group status for the selected participant samples.

Pilot threat condition 1. Next, for the pilot study, I used the FS dataset and subjected it to Threat Condition 1 (dataset FSC1). I followed this single condition through Phases II and II of the study to apply lessons-learned to the procedures for the remaining study conditions. For example, I used this process to create needed syntax files for analyses and to determine the most efficient sequence of analyses. I used the results to assist in the refinement of the covariate list for the corrective procedures as described below.

Refining covariate list. As the final part of Phase I, I identified a refined covariate list to use in the Phase IV adjustment techniques, e.g., constructing the propensity scores. While some variables were preliminarily identified in my proposal, Phase I provided an opportunity to explore the best and most parsimonious list of potential covariates to use in Phase IV analyses to maximize effectiveness. Justifying this in-depth consideration, Steiner et al. (2010) acknowledged that it is rarely possible to identify covariates meeting the strict requirements for propensity score analysis (PSA) in the planning phase of a study, meriting exploration during Phase I. They held that the importance of the covariate selection far outweighed the actual adjustment method ultimately used (Steiner et al., 2010). The determination of the covariates to use in PSA was decided in two ways: (a) developing a theory of selection (i.e., deciding what variables might be related to how cases were included in the two groups) and (b) using an iterative process of assessing the reduction of selection bias afforded by the adjustment (i.e., try different covariate combinations to form the propensity scores). Ultimately, the best covariate selection was based on those variables that were the most highly correlated with the selection process and the outcomes (Steiner et al., 2010) and those that provided the greatest bias reduction (Shadish & Steiner, 2010).

Phase II: Dataset Generation for Randomization Threat Conditions

The foundation datasets were manipulated in each condition to represent the specific threats to randomization. Realistic and plausible problem-scenarios guided the strategic exclusion of cases from the foundation datasets. This exclusion process was used to threaten the randomization balance of covariates between the intervention and control groups. To identify participant cases for *exclusion*, two different methods for each threat type were used to categorize or label cases according to the scenario context (described below). The purpose in using two different methods of strategic exclusion was to provide an alternative opportunity to confirm potential bias effects.

For the *allocation bias* scenario, I used the following as the basis for case exclusion: (a) baseline maternal risk index (summed composite of the presence of risk factors including teen mother, no high school, on public assistance, unemployed, and single mother--ranges from 0 to 5) and (b) baseline maternal welfare receipt (single covariate--ranges from 0 to 1). Maternal risk status was a computed variable already present in the EHS datasets as was the welfare variable.

For the *noncompliance* scenario, I used (a) a service intensity constructed variable present in the EHS datasets that was derived based on reported patterns of program participation at the 26-month interview (dichotomous variable indicating high or low intensity services based on a defined threshold--ranges 0 to 1), and (b) latent class analysis to identify cases' latent compliance classes (see Phase II analysis section for description).

For the *differential attrition* scenario: I used (a) a parent resources survey score reported at the 26-month interview (scores ranged from79 to 195) and (b) parents' average number of hours worked at the time of the 26-month interview (single score--ranges from 0 to 74).

Labeling Datasets for Conditions 1-27

A table follows each scenario description that shows each condition: the labels for the dataset threatened by the condition for both exclusion methods 1 and 2, a description of the condition, and the randomized comparison set to which the threatened set was compared. Again, FS, FM, and FL refer to the small foundation, medium foundation, and large foundation datasets. For example, FSC1 is the label for the data derived from the small foundation set subsequently exposed to Condition 1. The R label refers to the randomized data sampled from the foundation set to achieve the same sample size for equitable comparison. For example, FSC1-R is the label for the set randomly sampled from the small foundation dataset to equal that condition's sample size. The tables provide the labels and description for each of the datasets referred to in the procedures below.

Threat scenario 1: Allocation bias (conditions 1-9). This fictitious, yet plausible scenario involves a situation in which program staff are responsible for recruitment, enrollment, and subsequent randomization. They received instructions on how to randomly assign participants to groups but their implementation of the random assignment was insufficiently monitored by the researchers. The intended procedures involved staff opening the next envelope in an ordered stack of envelopes to reveal the assignment status for a newly enrolled participant. However, the assignment was insufficiently concealed and program staff were able to see through the envelopes and view the assignment printed on a paper inside. In one, two, or three sites (based on condition, sites were selected randomly), staff were not clear about the purpose of the random assignment or alternatively believed that random assignment was unethical due to withholding services from those most in need. After they examined the enrollment paperwork, they acted on their preference for serving the families with the highest level of risk. They selected envelopes out of sequence that ensured the families with the highest risk got into the EHS intervention group.

The end-goal for this scenario was that the highest risk families were no longer equally distributed across the intervention and control groups (as per the randomization) because staff placed a greater proportion of families with a higher risk composite in the intervention group. Since it was not possible to actually move highest risk families from the intervention group to the control group in the secondary dataset that was used, I created a targeted exclusion process to increase the proportion of higher risk families in the intervention group. In other words, I selected the specified proportion of the top highest-risk cases to exclude from the control group and I also selected an equal proportion of the bottom lowest-risk cases to remove from the intervention group. A multi-step process was used to create 18 datasets in this scenario (nine from each exclusion method).

***Step 1: Determine level of risk across all cases in the sample*.** As described above in the introductory paragraph for Phase II, the maternal risk index was used to identify the level of risk associated with each case. The distribution of high-risk families in the original sample was over 25%, enabling exclusion sampling as described below (ACF, 2002b).

Step 2a: Create conditions 1-3. This began with the randomly selected, small-sample foundation dataset (FS). In Condition 1, the top 12.5% of the highest risk cases were removed from the control group and the bottom 12.5% of the lowest risk cases were removed from the intervention group (new set FSC1). In Condition 2, the top 7.5% of the highest risk cases were removed from the control group and the bottom 7.5% of the lowest risk cases were removed from the intervention group (set FSC2). In Condition 3, the top 2.5% of the highest risk cases were removed from the control group and the bottom 2.5% of the lowest risk cases were removed from the intervention group (FSC3).

Step 2b: Create conditions 4-6. This step began with the randomly selected, medium-sample foundation dataset (FM). In Condition 4, the top 12.5% of the highest risk cases were removed from the control group and the bottom 12.5% of the lowest risk cases were removed from the intervention group (new set FMC4). In Condition 5, the top 7.5% of the highest risk cases were removed from the control group and the bottom 7.5% of the lowest risk cases were removed from the intervention group (set FMC5). In Condition 6, the top 2.5% of the highest risk cases were removed from the control group and the bottom 2.5% of the lowest risk cases were removed from the intervention group (FMC6).

Step 2c: Create conditions 7-9. This step began with the randomly selected, large-sample foundation dataset (FL). In Condition 7, the top 12.5% of the highest risk cases were removed from the control group and the bottom 12.5% of the lowest risk cases were removed from the intervention group (new set FLC7). In Condition 8, the top 7.5%

of the highest risk cases were removed from the control group and the bottom 7.5% of the lowest risk cases were removed from the intervention group (set FLC8). In Condition 9, the top 2.5% of the highest risk cases were removed from the control group and the bottom 2.5% of the lowest risk cases were removed from the intervention group (FLC9).

Step 3: Modify large sample foundation, medium sample foundation, and small sample foundation datasets to match sample sizes. In this step, I randomly deselected cases to create comparison datasets with equal sample sizes to the sets for Conditions 1-9. Note that as cases were selectively removed, the sample sizes decreased, thus the need for creating comparison groups that matched sample size.

Step 4: Conduct diagnostic and comparative analysis. This is described below in the analysis section and included examination of all four outcome variables for each threat condition.

Step 5: Repetition. Steps 2-4 were repeated using the single covariate method of exclusion (single risk factor of welfare receipt). Table 2 presents a summary of the nine conditions.

Table 2

Summary of Nine Conditions

	Set Labels			
Condition	Method 1	Method 2	Description	Randomized Comparison Set
1	FSC1	FSC1-2	allocation bias - small sample - large exposure	FSC1-R
2	FSC2	FSC2-2	allocation bias - small sample - medium exposure	FSC2-R
3	FSC3	FSC3-2	allocation bias - small sample - small exposure	FSC3-R
4	FMC4	FMC4-2	allocation bias - medium sample - large exposure	FMC4-R
5	FMC5	FMC5-2	allocation bias - medium sample – med. exposure	FMC5-R
6	FMC6	FMC6-2	allocation bias - medium sample - small exposure	FMC6-R
7	FLC7	FLC7-2	allocation bias - large sample - large exposure	FLC7-R
8	FLC8	FLC8-2	allocation bias - large sample – med. exposure	FLC8-R
9	FLC9	FLC9-2	allocation bias - large sample - small exposure	FLC9-R

Threat scenario 2: Early Head Start program noncompliance (Conditions 10-–18). In this scenario, as in the original sample, there were EHS intervention participants who do not participate in program activities at levels sufficiently believed to affect positive outcome change. While the reasons for nonparticipation were not important for this scenario per se, potential reasons might include high mobility or homelessness, for example. What was important was the pattern of EHS participation among individuals in the intervention group. The manipulation in this scenario involved increasing the proportion of noncompliers in the intervention group (i.e., those who participated in EHS activities the least often) by removing the designated proportions of EHS participants who were the most highly involved from the intervention group. Since in this scenario there was no variability in program use in the control group (i.e., should already equal zero participation in EHS), a random sample of participants was removed from the control group to match the sample size in the intervention group (i.e., controlling for potential sample size differences). As in Scenario 1, a similar multi-step process to create the datasets was subjected to the Scenario 2 threat conditions.

Step 1: Determine level of compliance across all participants in the sample. Again, the patterns of program use and involvement were thoroughly documented for the original sample. Recall that the involvement variables as measured in the 26-month parent services interview included (a) length of participation in EHS in months (divided into categories by quartile); (b) frequency of participation in EHS activities (categorical, numerous items specifying different activities); (c) time per month that parent spent in EHS activities (hours, reported in categories); and (d) time per month the focus child spent in EHS activities (hours, reported in categories) including child development services. A dichotomous service intensity variable was created based on these variables by MPR and is present in the original data (ACF, 2002b).

Step 2a: Create conditions 10-12. This began with the randomly selected, small-sample foundation dataset (FS). In Condition 10, the top 25% of the most highly involved/compliant cases were removed from the intervention group and 25% of the sample was randomly selected for removal from the control group (set FSC10). In Condition 11, the top 15% of the most highly involved/compliant cases were removed from the intervention group and 15% of the sample was randomly selected for removal from the control group (set FSC11). In Condition 12, the top 5% of the most highly involved/compliant cases was removed from the intervention group and 5% of the sample was randomly selected for removal from the control group (FSC12).

Step 2b: Create conditions 13-15. This step began with the randomly selected, medium-sample foundation dataset (FM). In Condition 13, the top 25% of the most highly involved/compliant cases were removed from the intervention group and 25% of the sample was randomly selected for removal from the control group (set FMC13). In Condition 14, the top 15% of the most highly involved/compliant cases were removed from the intervention group and 15% of the sample was randomly selected for removal from the control group (set FMC14). In Condition 15, the top 5% of the most highly involved/compliant cases were removed from the intervention group and 5% of the sample was randomly selected for removal from the control group (FMC15).

Step 2c: Create conditions 16-18. This step began with the randomly selected, large-sample foundation dataset (FL). In Condition 16, the top 25% of the most highly involved/compliant cases were removed from the intervention group and 25% of the sample was randomly selected for removal from the control group (set FLC16). In Condition 17, the top 15% of the most highly involved/compliant cases were removed from the intervention group and 15% of the sample was randomly selected for removal from the control group (set FLC17). In Condition 18, the top 5% of the most highly involved/compliant cases were removed from the intervention group and 5% of the sample was randomly selected for removal from the control group (FLC18).

Step 3: Modify large sample foundation, medium sample foundation, and small sample foundation datasets to match sample sizes. In this step, I randomly deselected cases to create comparison datasets with equal sample sizes to the sets for Conditions 10-18. Note that as cases were selectively removed, the sample sizes decreased, thus the need for creating comparison groups that matched sample size.

Step 4: Conduct diagnostic and comparative analysis. These procedures are described below in the data analysis section. This step included examination of all four outcome variables for each threat condition.

Step 5: Repetition. Steps 2-4 were repeated using the categorical and dichotomous variables (as required by the analysis; e.g., Eid, Langeheine, & Diener, 2003). The variables included participation variables for home visits, child care attendance, social events, and parenting classes. Latent class analysis was used to determine the class or group to which each participant belonged, representing their latent participation pattern. This method is described below in the data analysis section. Table 3 presents a summary of Conditions 10-18.

Table 3

Summary of Conditions 10-18

	Set Labels			Randomized
	Method	Method		Comparison
Condition	1	2	Description	Set
10	FSC10	FSC10-2	noncompliance bias-small sample-large exposure	FSC10-R
11	FSC11	FSC11-2	noncompliance bias-small sample-medium exp.	FSC11-R
12	FSC12	FSC12-2	noncompliance bias-small sample-small exposure	FSC12-R
13	FMC13	FMC13-2	noncompliance bias-medium sample-large exp.	FMC13-R
14	FMC14	FMC14-2	noncompliance bias-medium sample-medium exp.	FMC14-R
15	FMC15	FMC15-2	noncompliance bias-medium sample-small exp.	FMC15-R
16	FLC16	FLC16-2	noncompliance bias-large sample-large exposure	FLC16-R
17	FLC17	FLC17-2	noncompliance bias-large sample-med. exposure	FLC17-R
18	FLC18	FLC18-2	noncompliance bias-large sample-small exposure	FLC18-R

Threat scenario 3: Differential attrition (conditions 19-27). In this scenario, it was supposed that Early Head Start (EHS) program families who had stronger employment and overall resources left the EHS study before the 36 months data collection point because they no longer believed they needed EHS services. Guo and Fraser (2010) indicated that participants who were more likely to drop out were often the ones who perceived they no longer received benefits from the intervention (p. 280). Thus, there becomes a missing-not-at-random (MNAR) problem. The goal in this scenario was to decrease the proportion of the sample in the intervention group at 36 months who were employed and had higher resources. It was presumed they were too busy to participate in EHS home visits because they were working. They could now afford other child care situations; thus, they dropped out of the research and the EHS program. Because they were not involved in the EHS services or home visits, the control group attrition occurred at random. Thus, a random proportion was removed from the control group.

Step 1: Determine level of resource across all participants in the sample. The variables used to create this composite variable included employment, high school diploma, and the reported adequacy-of-resources variables from the 26-month parent services interview. The interview examined the adequacy of resources including food, housing, money, medical care, transportation, child care, support from friends, and parent information. The summed composite of the presence of employment, high school diploma, and the nine resource variables ranged from 0 to 11.

Step 2a: Create conditions 19-21. This step began with the randomly selected, small-sample foundation dataset (FS). In Condition 19, the top 25% of the most highly resourced cases were deselected from the intervention group and 25% of the sample was randomly deselected from the control group (set FSC19). Note that the term deselected was used rather than removed, as in previous scenarios, because these excluded cases (minus their 36-month outcome data) needed to remain in the dataset for use in Phase II and IV analyses. In Condition 20, the top 15% of the most highly resourced cases was deselected from the intervention group and 15% of the sample was randomly deselected from the control group (set FSC20). In Condition 21, the top 5% of the most highly resourced cases was deselected from the intervention group and 5% of the sample was randomly deselected from the control group (FSC21).

Step 2b: Create conditions 22-24. This step began with the randomly selected, medium-sample foundation dataset (FM). In Condition 22, the top 25% of the most highly resourced cases was deselected from the intervention group and 25% of the sample was randomly deselected from the control group (set FMC22). In Condition 23, the top 15% of the most highly resourced cases was deselected from the intervention group and 15% of the sample was randomly deselected from the control group (set FMC23). In Condition 24, the top 5% of the most highly resourced cases was deselected from the intervention group and 5% of the sample was randomly deselected from the control group (FMC24).

Step 2c: Create conditions 25-27. This step began with the randomly selected, large-sample foundation dataset (FL). In Condition 25, the top 25% of the most highly resourced cases was deselected from the intervention group and 25% of the sample was randomly deselected from the control group (set FLC25). In Condition 26, the top 15% of the most highly resourced cases was deselected from the intervention group and 15% of the sample was randomly deselected from the control group (set FLC26). In Condition 27, the top 5% of the most highly resourced cases was deselected from the intervention group and 5% of the sample was randomly deselected from the control group (FLC27).

Step 3: Modify large sample foundation, medium sample foundation, and small sample foundation datasets to match sample sizes. In this step, I randomly deselected cases to create comparison datasets with equal sample sizes to the sets for Conditions 19–27. Note that as cases were selectively excluded, the sample sizes decreased, thus the need for creating comparison groups that matched sample size.

Step 4: Conduct diagnostic and comparative analysis. These procedures are described below in the data analysis section. This included examination of all four outcome variables for each threat condition.

Step 5: Repetition. Steps 2-4 were repeated using a single score for parent resource, i.e., the number of hours worked at the time of the 26-month interview (single score ranged from 0 to 1), excluding cases on the basis of greater hours worked (see Table 4 for summary of conditions 19-27).

Table 4

Summary of Conditions 19-27

	Set Labels			Randomized
Condition	Method 1	Method 2	Description	Comparison Set
19	FSC19	FSC19-2	diff attrition bias-small sample-large exposure	FSC19-R
20	FSC20	FSC20-2	diff attrition bias-small sample-medium exp.	FSC20-R
21	FSC21	FSC21-2	diff attrition bias-small sample-small exposure	FSC21-R
22	FMC22	FMC22-2	diff attrition bias-medium sample-large exp.	FMC22-R
23	FMC23	FMC23-2	diff attrition bias-medium sample-medium exp.	FMC23-R
24	FMC24	FMC24-2	diff attrition bias-medium sample-small exp.	FMC24-R
25	FLC25	FLC25-2	diff attrition bias-large sample-large exposure	FLC25-R
26	FLC26	FLC26-2	diff attrition bias-large sample-med. exposure	FLC26-R
27	FLC27	FLC27-2	diff attrition bias-large sample-small exposure	FLC27-R

Table 5 summarizes each scenario, the manipulation performed for intervention and control groups (exclusion factors), and the hypothesized effects on the outcomes.

Table 5

Summary of Threat Scenarios and Manipulations by Group

Group	Scenario 1: Biased Allocation	Scenario 2: EHS Noncompliance	Scenario 3: Differential Attrition
Premise	Increase proportion of high risk cases in intervention group	Increase proportion of cases less intensely involved in EHS services	Decrease proportion of cases in the intervention group who are highly resourced
Manipulation to Threatened Intervention Group	Remove cases with lower risk composite scores	Remove cases with highest intensity of EHS service involvement	Remove cases with the lowest resource scores
Manipulation to Threatened Control Group	Remove cases with higher risk scores	Remove a randomly selected sample to match sample size in intervention group	1.Remove a randomly selected sample to match sample size in intervention group 2. Retain same sample size in control group*
Randomized Comparison Group	Use random samples from foundation sets to match on sample size to isolate bias not power effects.	Use random samples from foundation sets to match on sample size to isolate bias not power effects.	1. Use random samples from foundation sets to match on sample size to isolate bias not power effects. 2. Use actual foundation sets (FL, FM, FS) to more realistically mimic attrition from intervention only (sample size will be larger for the randomized comparison)

* For consistent methodology, I created equal sample sizes for intervention and control to isolate bias versus power effects.

** If attrition were occurring in the intervention sample only, as proposed here, then the sample size for the control would remain unchanged. This is part of the problem with differential attrition (McKnight et al., 2007).

Phase III and IV Procedures

Since Phases III and IV are fully data analytic phases, they are described below in the next section.

Data Analysis

Numerous analytic tools and strategies were employed in the different phases of the study. Several statistical software packages and data management tools were used to perform the analyses: Excel (for organizing data outputs, conducting simple calculations, and creating charts and tables); SPSS version 18 (to conduct descriptive, univariate, and multivariate techniques including chi-square tests, *t*-tests, repeated measures ANOVA, and logistic regression); and M*plus* version 5 (Muthén & Muthén, 1998–2011) for the latent variables.

The following research questions guided the study:

Q1 What are the comparative effects of 27 randomization threat conditions on the randomization of the EHSRE data?

a. What evidence of intervention-control group covariate imbalance (i.e., baseline inequivalence between intervention and control groups) is revealed in each of 27 threat conditions?

b. What is the level of bias introduced by each of 27 threat conditions?

Q2 How sensitive are EHSRE study results to randomization threat conditions that include manipulations of threat type (i.e., biased allocation, noncompliance, and differential attrition), overall sample size, and proportion of sample exposed to the threat condition? Specifically,

a. To what extent do threatened and non-threatened samples differ on child outcome scores and observed effect sizes?

b. What is the rate of Type I and Type II error associated with the threatened samples compared with the associated non-threatened sample for each of the child outcome variables?

Q3 To what degree are corrective statistical methods effective in restoring or distorting results in the face of typical randomization threat conditions? Specifically,

a. What effect do corrective methods have on findings (means, significance, and effect sizes) from threatened samples?

b. Within threat types, what is the comparative effectiveness of two corrective techniques in reducing bias introduced under threat conditions?

Phase I Analysis

To gain a clearer understanding of the variables included in the Early Head Start (EHS) databases, I used SPSS to conduct descriptive and frequency analyses in Phase I to examine the means, standard deviations, ranges, frequencies, and distributions of the variables of interest (i.e., covariates and outcomes). I also examined the relationships among the variables selected for the study by conducting bivariate Pearson correlations. In addition, further investigation of the relationships of the covariates to the outcome variables was conducted using a series of analyses in SPSS to determine which covariates to use in each of the corrective methods. Each of the above analyses took place using the unthreatened original EHSRE data. The analyses for the pilot study underwent the appropriate steps described in Chapter IV.

Phase II Analysis

Since Phase II involved the creation of a variety of datasets through sampling, exploratory analyses were used to verify that the foundation datasets retained the balance achieved through randomization. The statistical analyses I conducted to check the randomization, i.e., verifying the similarity of the intervention and control groups, included (a) independent samples *t*-tests to compare variable means between the intervention and control groups for the continuous variables and (b) chi-square tests of independence to compare distributions of categorical variables between the intervention and control groups. These were tested to determine if the difference was significantly different at the .10 level using a two-tailed test. The use of these analyses and this significance level was selected based on the precedent set by MPR for the original EHSRE (ACF, 2002b). By remaining consistent with the original method, I generally referred to differences or similarities between the present study's findings and the original.

Each of the datasets subsequently constructed from the foundation sets to represent various threat conditions were also analyzed as described in the paragraph above to address Research Question 1a. To address Research Question 1b, single-sample *t*-tests and chi-square tests were used to compare threatened means and frequencies to randomized means and frequencies (see Table 6).

Table 6

Addressing Research Questions 1a and 1b

Research Question	Assessing	Comparison Type	Analyses	Criteria
1a	imbalance	intervention v. control	independent samples *t*-tests and chi-square tests of baseline characteristics and outcomes	sig difference at α = .1, two tailed[1] and no more than 10% of tests may be significantly different
1b	bias	threatened v. randomized	single-sample *t*-tests and chi-square tests of baseline characteristics and outcomes for continuous variables	sig difference at α = .1, two tailed

[1] Precedent set in EHSRE analysis (ACF, 2002b).

Assessing Compliance Type Using Latent Class Analysis

To identify the proportions of intervention cases (in the nine noncompliance datasets) by latent compliance type so I could strategically exclude proportions of strong compliers, I used M*plus* version 5 to conduct latent class analysis (LCA) and guidance using Hagenaars and McCutcheon's (2002) text, *Applied Latent Class Analysis*. The purpose of the LCA was to (a) create a model that permitted categorization of intervention participants into different types/classes of Early Head Start (EHS) program compliers such as non and weak compliers, moderate compliers, and strong compliers; (b) confirm how many latent classes of compliers best fit the data; (c) categorize cases by compliance type; and (d) identify how many cases (i.e., the proportion) would be considered strong compliers. Two relevant hypotheses guided this work. First, from the Raikes et al. (2006) study, I expected that there would be four latent classes of compliance types, justifying the use of this confirmatory procedure. In actuality, the best model fit was with five classes (see Chapter IV for results). Then, I hypothesized that the proportion of strong compliers would exceed 25%, thus enabling the exclusion of the specified proportions for the conditions based on evidence in the ACF (2002b) reports; this was indeed the case. The analyses involved the following steps:

Step 1: Data preparation. The Statistical Package for the Social Sciences (SPSS) was used to output an ASCII version of the data for input into M*plus.* Data were verified to ensure that the input file was successfully formed.

Step 2: Model specification. Using the variables identified in Step 1 of the description of Threat Scenario 2, I specified models in M*plus.* The variables describing service use were used to cluster cases around their shared variability. The method of estimation I selected was the MLR estimator, which is an iterative maximum likelihood (robust) estimator used with an expectation-maximization (EM) algorithm, commonly used in LCA per Hagenaars and McCutcheon (2002). In M*plus,* the variable names, the number of classes (four), the analysis type, desired plots, and outputs are specified.

Step 3: Model evaluation. The outputs showed the classification of cases based on their most likely class membership. The output also provided the number of cases in each class; it indicated the means for each of the input variables and how they combined to form/describe the classes (Muthén & Muthén, 1998–2011). Statistical tests for evaluating the fit of a latent class model are based on a comparison of the observed frequencies of the response patterns and the frequencies of the response patterns expected on the basis of a latent class model (Eid et al., 2003). I assessed goodness-of-fit and reviewed information criteria. Criteria for model evaluation included the Pearson chi-square, the likelihood ratio chi-square, the Akaike information criteria, and the Bayesian information criteria. The Pearson chi-square is a measure of the comparison of the actual response patterns (of variables in the model) with what is expected under the model. If the model has a low *p* value, then it is does not have good fit. Lower likelihood ratio chi-square, Akaike information criteria, and Bayesian information criteria values suggest better fit. Akaike information criteria and Bayesian information criteria are model comparison measures and are used to decide how many classes provide the best fit for the data. These measures were taken together to decide which model and number of classes was best (Garraza, Azur, Stephens, & Walrath, 2010; Hagenaars & McCutcheon, 2002).

Step 4: Determine number of classes/compliance types. After the model was confirmed, I used the proportion of cases that fit into the strongest compliance types to form the basis of the exclusion criteria for data generation in Phase II. Given the patterns of compliance evident in the EHSRE study, I hypothesized that a strong complier class would emerge from the data and this was true.

Phase III Analysis: Sensitivity Analysis

Sensitivity analyses allow a researcher to assess the impact that changes across study conditions or parameters have on the study results and conclusions. They are often used in simulation studies, in econometrics, or in medical research to describe the effects of different scenarios on the outcomes of interest. Taylor (2009) described the simplest form of sensitivity analysis is to simply vary one value in the model by a given amount and examine the impact the change has on the model's results. One-way sensitivity analysis is when one parameter is changed at a time, i.e., the changes from 5% to 15% in the proportion of sample affected by the threat condition. The sensitivity analyses included comparative examination of the main effects of intervention group differences and change-over-time on each of the four outcome variables across each of the 27 conditions (each created using two different methods) to determine if results differed between the randomized data set and threatened datasets.

As a first step, routine analyses were conducted to examine changes in outcomes over time and to determine differences between treatment and control in threatened and non-threatened, randomized sets. Point-in-time group comparisons (using data collected at 24 and 36 months), change over time (within and between for all points available in the dataset), and estimates of the intervention effects (level of significance and effect sizes) were also examined. Point-in-time comparisons were conducted using independent samples *t*-tests to compare intervention and control groups on each of the estimated intervention effects of the four outcomes measures at 24 and 36 months. These tests were conducted for both the sample-under-threat and the randomized sample. The PPVT was only collected at 36 months so was not included in the change over time analyses. The CBCL, BSID-II, and the Engagement of Parent variables were examined for change over time using repeated measures ANOVA (see Table 7 for a summary of the routine analyses and sensitivity analyses). At the second stage of analysis to address Research Questions 2a and 2b regarding the sensitivity analyses, mean and effect size differences were compared for threatened and non-threatened samples. To assess mean differences, a percent bias estimate was computed, indicating the difference between the outcome mean under threat condition and the expected mean from the randomized comparison using the formula below:

$$B(\bar{y}_c) = ((\bar{y}_c - \bar{y}_r) / \bar{y}_{r)} * 100\%,$$

where $\bar{y}_c$ is the mean outcome score for the sample under a threat condition and $\bar{y}_r$ is the mean outcome score for the randomized comparison sample. I also assessed the rate of Type I and Type II error in the threatened samples. Thus, I was interested in examining whether the threatened samples would falsely find statistically significant effects or whether they would falsely reveal no statistically significant findings.

For effect size differences, I examined the effect size confidence intervals and the magnitude of the effect size differences between threatened and randomized findings. Table 8 displays the comparison and associated criteria. Ultimately, these various pieces of evidence were intended to be taken together to assess whether different conclusions would be drawn from threatened samples compared with the randomized samples. Chapter IV results provide both individual condition and summarized results on each of these indicators, a judgment of overall bias derived from the findings-as-whole, and the likelihood of a different conclusion being drawn.

Table 7

Addressing Research Questions 2a and 2b

Research Question	Step 1 Outcome Analyses	Step 1 Comparison Assesses	Step 2 Sensitivity Analyses	Step 2 Assesses
2a and 2b	1) point-in-time analysis of outcome variables using independent *t*-tests 2)Change over time repeated measures ANOVA	1) statistically sig differences between intervention and control 2) statistically sig differences over time within groups and between intervention and control	1&2) comparison of effect size estimate confidence intervals[1] and magnitude of effect size differences[2]; comparison of means using percent bias[3]; and assessment of type I and II error in threatened samples	1&2) for threatened v. randomized: differences in effect sizes and differences in findings of significance (Type I and Type II error rate among threatened samples); and overall percent bias introduced by the threat condition

Note: All tests of significance were conducted at α = .05, two tailed

[1] Examination of confidence intervals was in terms of overlap; non-overlapping CIs are consistent with effect size differences (Thompson, 2007).

[2] Any effect size differences greater than .20 was considered analogous to a small effect (Cohen, 1988), thus would effectually result in a different conclusion being drawn.

[3] The percent bias criterion used to indicate meaningful bias was any percent bias greater than 5% (using a 95% reference) for a value greater than the expected proportion of values different from the mean (Johnson, 2008).

Phase IV Analysis: Corrective Techniques

For each of the threat types, two corrective techniques were employed to attempt adjustments when bias was introduced to the findings by the threat conditions. The adjustment techniques used in the study were propensity score analysis and analysis of covariance. The corrective techniques were also applied when bias was not introduced to assess the overall effects of unneeded adjustment and to mimic what might occur in practice. The comparison of findings across threatened and non-threatened samples within threat types was intended to assess whether a given adjustment technique improved the accuracy of the findings (i.e., after adjustment, do the findings more closely resemble the randomized outcome mean, significance, effect size) or whether the correction further distorted the findings (i.e., greater difference from the randomized outcome means, significance, effect size). I also examined the percentage of bias reduction (or increase) found upon comparison of the corrected/adjusted outcome analyses with those analyses from the randomized samples. Since this was not a crossed design, meaning that not every corrective technique was applied to every threat type, comparisons of adjustment effectiveness within, not across, threat types were made. I compared the relative efficacy between the two techniques (within, not across threat types) in the amount of bias reduction realized.

Propensity score analysis using propensity scores as weights was used to attempt adjustment across all three threat types. The second adjustment used was a simple covariate adjustment (ANCOVA) using variables associated with the outcome variables. These variables were selected because they were continuous variables associated with differences in the outcome variables. In the case of allocation bias, the mother’s age was used as the covariate. For noncompliance and differential attrition adjustment procedures, I used the total number of hours the parent worked as the covariate.

Following the analyses to apply the corrective techniques, the same patterns of outcome and sensitivity analyses were conducted as were described for Research Question 2 to address Research Question 3; however, the results were compared across randomized, threatened and corrected samples to gauge whether the corrected/adjusted findings improved the estimates or, in other words, if the results resembled the randomized data more closely than the threatened data.

Simple covariate adjustment. The most common method of handling imbalance between intervention and control groups is to include variables for which there are significant differences between the two groups as covariates (Taylor & Innocenti, 1993). For this analysis, I used ANCOVA (after checking appropriate assumptions) to control for baseline differences to obtain new estimates of the intervention effects (means, significance level, and effect size) for all the aforementioned comparisons.

Propensity score analysis. Several resources supported the steps and procedures needed to construct propensity scores and to use them in subsequent analyses to adjust for covariate imbalance between the intervention and control groups. In particular, the following resources were particularly helpful: *A Primer on Propensity Score Analysis* (Shadish & Steiner, 2010), *Some Practical Guidance for the Implementation of Propensity Score Matching* (Caliendo & Kopeinig, 2008), *Propensity Score Analysis and its Applications* (workshop materials from Guo & Fraser, 2010), and *Practical Applications of Propensity Scores* (Clark, 2008).

Step 1: Covariate selection. The baseline variables selected to construct the propensity scores were limited to the variables collected as part of the EHSRE study; however, as alluded to earlier, there was a substantial number from which to choose from the baseline data collection prior to randomization. Otherwise, theoretically, the number of covariates that could be included in the model was not limited in the literature for parsimony's sake. The covariates selected as part of the Phase I investigation of these variables helped identify variables that were related to selection into the intervention and related to the outcome.

***Step 2: Logistic regression to obtain propensity scores*.** Propensity scores were estimated using the covariate set to predict each participant case's likelihood of selection into the intervention group using logistic regression in SPSS. The predicted probabilities were saved in the dataset to use for diagnostics and subsequent analyses. Model adequacy was checked, for example, through examining the equivalence of the distributions of the propensity scores in the intervention and control groups, as well as calculating the ratio of the intervention and control variances of the propensity scores (ratios close to 1; Shadish & Steiner, 2010). In addition, baseline covariate balance was checked again using independent *t*-tests.

Step 3: Weighting and outcome analysis. In SPSS, the propensity scores were applied as weights (using the weighting function in SPSS to adjust the analysis) and the same outcome and sensitivity analyses described for Phase III were repeated.

***Step 4: Repetition*.** Steps 1 through 3 were repeated for each of the threatened conditions.

Step 5: Conduct sensitivity analysis. To assess the degree to which the propensity score analysis assisted or thwarted attempts to correct imbalance and to improve outcome estimates, I conducted the same set of sensitivity analyses as were performed for the threatened versus randomized outcomes. These analyses included calculation of effect size differences, examination of Type I and Type II errors, and percent mean bias. These analyses were compared across the threatened, adjusted, and randomized findings to determine relative effects.

Propensity score analysis to adjust for differential attrition bias. As an alternative adjustment approach, propensity score analysis can be used to reduce selection bias due to differential attrition. This procedure can rebalance the intervention and control groups to restore equivalence; however, it does not replace missing data. Following the steps described to adjust for allocation bias, propensity score analyses were used and examined for their comparative effectiveness in restoring results achieved through randomization.

Summary

Phases I through IV of the current study were conducted using the procedures as specified. To reiterate, Phase I was a data preparation and pilot phase and Phase II comprised the datasets generation for the 27 threat conditions (two exclusion methods each for a total of 54 threatened datasets); analyses from the threat conditions were used to answer Research Question 1. Phase III examined sensitivity of the EHS study results to each of the threat conditions to respond to Research Question 2 and Phase IV explored the effectiveness of the corrective methods in restoring or distorting the data under threat conditions to address Research Question 3. In the results section (see Chapter IV), I provide narrative and tabular results to organize and synthesize the study findings.

CHAPTER IV

RESULTS

Phase I: Data Merging and Pilot Analyses

The initial phase of the analysis involved the development of the original base set (OBS), which was derived by gathering the necessary variables from each of the original Early Head Start databases and combining them into a single database for use in this study. A total of 94 variables and 1,756 cases were the results of this effort. This database was subsequently randomly subsampled to create three foundation datasets of large, medium, and small sample sizes. The final sample sizes, prior to Phase II analyses, were as follows: (a) foundation large, N = 1,400 (700 control and 700 program); (b) foundation medium, N = 600 (300 control and 300 program); and (c) foundation small, N = 250 (125 control and 125 program). After creating the foundation sets, independent samples *t*-tests (using $\alpha = .10$) were performed to examine whether the subsampling process affected the intervention-control balance on 12 baseline characteristics. Randomization was retained for each of the foundation datasets in terms of their baseline characteristics.

The pilot study, involving a full run-through of Condition 1 (FSC1; small sample, high proportion [25%] affected), resulted in the development of a process for strategically excluding cases using SPSS. Condition 1 (with exclusion method 1) involved using the categorical maternal risk composite variable to identify and randomly deselect 12.5% of cases from the small foundation dataset with a high level of risk from the control group and to randomly deselect 12.5% of cases from the same dataset with a low level of risk from the intervention group. Another new dataset was created to serve as the randomized comparison by randomly deselecting cases from the same small foundation dataset to match the intervention and control sample sizes in the FSC1 sample (see descriptive statistics in Tables 8-19).

For the pilot analysis that correspond to the later Phase II analyses for the other conditions, the balance checks of the data (using independent samples *t*-tests and chi-square tests of independence) indicated the control and intervention groups were statistically significantly different for 5 of the12 baseline characteristic, showing that imbalance was created. The variables that differed between intervention and control in the threatened sample were parent age, education, primary occupation, living arrangement, and maternal risk. Which variables were different, however, was immaterial for the purpose of this analysis--to determine *how many* variables of the 12 were out of balance. If more than one was out of balance, then this was considered out of balance.

Three tests indicated statistically significant differences between the threatened and randomized comparison sets using one-sample *t*-tests with the randomized dataset means used as the target means (education, living arrangement, and maternal risk). Similar to the balance tests, the quantity, rather than the variables themselves, was of interest; thus, if more than one test was determined to indicate bias between threatened and randomized, then threat condition was determined to have introduced biased. Thus, this analysis showed that bias was created in the FSC1 threatened condition (see Tables 8-19 for the pilot's detailed baseline balance and bias analytic results).

Table 8 shows the means and standard deviations of the threatened and randomized sets for the pilot. This is one example of the descriptive analyses that were performed for each of the threat conditions and corresponding randomized datasets.

Table 8

Pilot Means and Standard Deviations by Group for Foundation Small Condition 1 and Foundation Small Condition 1-Randomized on Five Continuous Baseline Characteristics

		FSC1			FSC1-R		
	Group	*n*	Mean	*SD*	*n*	Mean	*SD*
Mother Age	C	93	23.409	5.634	93	22.355	5.183
	I	93	21.763	6.478	94	22.872	6.193
Child Gender	C	93	.505	.503	93	.559	.499
	I	94	.596	.493	94	.543	.501
Child Premature	C	62	.129	.338	65	.077	.269
	I	68	.162	.371	65	.154	.364
Welfare Receipt	C	76	.250	.436	76	.276	.450
	I	71	.282	.453	69	.188	.394
Food Stamps Receipt	C	86	.337	.476	87	.414	.495
	I	89	.427	.497	89	.292	.457

Key: C = control group; I = intervention group; FSC1 refers to the "foundation small condition 1" threat condition, and FSC1-R refers to the randomized comparison set.

Slight mean differences are observed between threatened and randomized data on each of the variables. These were tested for statistical significance with results shown in Table 16. While means were computed for the five continuous variables presented in

Table 8, Tables 9 through 15 display the frequencies for the seven categorical baseline variables. The pilot's descriptive and frequency tables are examples of the types of tables that were calculated for every threat condition (created using method 1 and method 2) and for every corresponding randomized set. Note that later in the Phase II through Phase IV results, these detailed tables were not presented because there are several hundred pages of results. I have instead selected examples to display in Appendices C through E. Summary tables, however, are presented that show the overall results aggregated in different ways over conditions.

Table 9 shows frequencies on the counts for the categorical variable of race. Because the sample sizes are equivalent, the counts might be directly compared in each cell. Slightly different numbers of participants by race are found in the threatened condition versus the randomized data. Results of the statistical significance tests of the differences in frequencies of race across the cell (Chi-square tests) are presented in Table 17, as are the findings for the next five categorical variables.

Table 9

Pilot Frequencies by Experimental Group for Foundation Small Condition 1(Method 1) and Foundation Small Condition 1-Randomized for Race

				FSC1					FSC1-R		
Race	Group	A	B	C	O	Total	A	B	C	O	Total
	C	37	25	24	5	91	43	27	17	4	91
	I	32	31	22	3	88	35	27	23	2	87
	Total	69	56	46	8	179	78	54	40	6	178

Key: A = White; B = African American; C = Hispanic/Latino; O = Other Race; C = control group; I = intervention group.

Table 10

Pilot Frequencies by Experimental Group for Foundation Small Condition 1 (Method 1) and Foundation Small Condition 1-Randomized for Education

		FSC1				FSC1-R			
Education	Group	A	B	C	Total	A	B	C	Total
	C	26	30	30	86	34	28	25	87
	I	50	25	11	86	36	29	23	88
	Total	76	55	41	172	70	57	48	175

Key: A = Less than 12th grade; B = High School Diploma or GED; C = More than High School Education; C = control group; I = intervention group.

Table 11

Pilot Frequencies by Experimental Group for Foundation Small Condition 1 (Method 1) and Foundation Small Condition 1-Randomized for Primary Occupation

		FSC1				FSC1-R			
Primary Occupation	Group	A	B	O	Total	A	B	O	Total
	C	31	12	42	85	26	16	45	87
	I	14	22	50	86	17	20	50	87
	Total	45	34	92	171	43	36	95	174

Key: A = Employed; B = School or Training; O = Other primary occupation; C=control group; I = intervention group.

Table 12

Pilot Frequencies by Experimental Group for Foundation Small Condition 1 (Method 1) and Foundation Small Condition 1-Randomized for English Language Ability

		FSC1				FSC1-R			
Language Ability	Group	A	B	C	Total	A	B	C	Total
	C	71	2	14	87	76	2	10	88
	I	70	6	9	85	70	7	9	86
	Total	141	8	23	172	146	9	19	174

Key: A = Parent's primary language is English; B = Primary language is not English but the parent speaks English well; C = The primary language is not English and the parent does not speak English well; C = control group; I = intervention group.

Table 13

Pilot Frequencies by Experimental Group for Foundation Small Condition 1 (Method 1) and Foundation Small Condition 1-Randomized for Living Arrangement

		FSC1				FSC1-R			
Living Arrange-ment	Group	A	B	C	Total	A	B	C	Total
	C	34	28	31	93	25	34	34	93
	I	19	46	29	94	28	39	27	94
	Total	53	74	60	187	53	73	61	187

Key: A = Lives with husband; B = Lives with Other Adults; C = Lives alone; C = control group; I = intervention group.

Table 14

Pilot Frequencies by Experimental Group for Foundation Small Condition 1 (Method 1) and Foundation Small Condition 1-Randomized for Maternal Risk Index

		FSC1				FSC1-R			
Maternal Risk	Group	A	B	C	Total	A	B	C	Total
	C	51	27	0	78	37	30	15	82
	I	16	31	31	78	39	22	21	82
	Total	67	58	31	156	76	52	36	164

Key: A = 0, 1, or 2 risks; B= 3 risks; C = 4, 5 risks; C = control group; I = intervention group.

Table 15

Pilot Frequencies by Experimental Group for Foundation Small Condition 1 (Method 1) and Foundation Small Condition 1-Randomized for Child Age

		FSC1				FSC1-R			
Child Age	Group	A	B	C	Total	A	B	C	Total
	C	22	33	38	93	21	32	40	93
	I	22	37	35	94	25	35	34	94
	Total	44	70	73	187	46	67	74	187

Note: Child's age at EHS application; A = Mother Pregnant; B = Child Less than 5 months old; C = Child More than 5 months old; C=control group; I=intervention group.

Tables 16 and 17 show the pilot results of the balance check comparing the equivalence of the baseline characteristics between intervention and control. What to look for is the *introduction of imbalance* (i.e., more than one statistically significantly different test of the 12 tests) in the threatened condition (here labeled FSC1) and the *retention of balance* in the randomized data on the same baseline characteristics (i.e., no more than one statistically significantly different test of the 12 tests). These tests were conducted at the 10% level to match methods from the EHSRE study.

Note in Table 16 for FSC1, one of the baseline characteristics is statistically significantly different between intervention and control groups and similarly for FSC1-R. Then, when combined with the results in Table 17, FSC1 has four more statistically significantly different intervention versus control tests to make a total of 5 of the 12 baseline characteristics that are different. Thus, imbalance was introduced in the threatened condition. The randomized set, with its reduced sample size, retained balance with only the one significant test.

Table 16

Foundation Small Condition 1 and Foundation Small Condition 1-Randomized Pilot Balance Check: Independent Samples t-Tests for Continuous Characteristics

	FSC1			FSC1-R		
	t	df	Sig. (2-tailed)	*t*	df	Sig. (2-tailed)
Mother's Age	1.848	184	.066*	-.619	185	.536
Child Gender	-1.241	185	.216	.227	185	.821
Child Premature	-.524	128	.601	-1.372	117	.173
Welfare	-.432	145	.666	1.254	142	.212
Food Stamps	-1.219	173	.224	1.692	172	.092*

Note: Asterisk (*) indicates statistically significant difference at alpha = .10.

Table 17

Foundation Small Condition 1 and Foundation Small Condition 1-Randomized Pilot Balance Check: Chi-squares Test of Independence for Categorical Characteristics

	FSC1			FSC1-R		
	X^2	df	Sig. (2-sided)	X^2	df	Sig. (2-sided)
Race	1.542	3	.673	.230	3	.973
Education	16.838	2	<.001*	.152	2	.927
Primary Occupation	10.054	2	.007*	2.591	2	.274
Language Ability	3.071	2	.215	3.054	2	.217
Living Arrangement	8.685	2	.013*	1.310	2	.519
Maternal Risk Index	49.559	2	<.001*	2.283	2	.319
Child's Age	.347	2	.841	.963	2	.618

Note: Asterisk (*) indicates statistically significant difference at alpha = .10.

Tables 18 and 19 combined demonstrate that bias was introduced as a result of the threat exclusion process for FSC1. Three of the 12 significance tests showed differences between threatened and randomized data corresponding to those three variables (recall criteria that no more than one may be significantly different to remain unbiased on baseline characteristics).

Table 18

Foundation Small Condition 1 Pilot Bias Check for Continuous Variables (Comparing Disrupted to Randomized Standard, One Samples t-Tests)

	FSC1					
	N	Mean	*SD*	*t*	df	sig.
Mother's Age	186	22.586	6.110	-0.065	185	.948
Child Gender	187	0.551	0.499	0.000	186	1.000
Child Premature	130	0.146	0.355	0.989	129	.325
Welfare	147	0.265	0.443	0.843	146	.401
Food Stamps	175	0.383	0.487	0.829	174	.408

None of the continuous baseline variable means differed between threatened and randomized datasets (i.e., used the randomized mean as the target for the threatened mean in one-samples *t*-test). However, in Table 19, the categorical variables of parent education, living arrangement, and maternal risk are statistically significantly different between threatened and randomized, thus indicating bias introduced in the threat condition. The direction of the difference was not hypothesized; the variable itself had limited relevance for the purpose of simply assessing the *number* of tests that showed a difference.

Table 19

Foundation Small Condition Pilot Phase II Bias Check for Categorical Variables

	FSC1			
	N	X^2	df	Sig. (2-sided)
Race	179	3.016	3	.389
Education	171	15.282	2	<.001*
Primary Occupation	171	2.891	2	.236
Language Ability	172	2.072	2	.355
Living Arrangement	187	8.861	2	.012*
Maternal Risk Index	156	42.605	2	<.001*
Child's Age	187	0.683	2	.711

Note: Asterisk (*) indicates statistically significant difference at alpha = .10.

While the previous pilot tables showed examples of assessment of baseline balance and bias that were found in Phase II, the next set of pilot analyses and tables exemplify what was found in Phase III. Recall that Phase III was designed to conduct outcome and sensitivity analysis with the threatened and randomized datasets.

In the pilot's analyses using FSC1 (method 1) and FSC1-R, the results between the two sets differed on a few outcomes, namely child engagement of parent at both 24 months and 36 months. The subsequent sensitivity analyses were used to assess whether the outcome results using the threatened dataset differed meaningfully (recall criteria from Tables 6 and 7) from those obtained using the matching randomized dataset. Table 20 shows that the effect sizes for child engagement were not only different by about a medium effect size but were in the opposite direction. For child engagement at 24 months, a false statistically significant difference between intervention and control was found (i.e., compared to no differences found in the randomized data).

Table 20

Foundation Small Condition 1 Pilot: Intervention vs. Control Point-in-Time Child Outcome Analyses

	Threatened Dataset				Randomized Dataset			
	Test Value (*t*)	df	Sig Level α=.05	Effect Size (*d*)	Test Value (*t*)	df	Sig Level α=.05	Effect Size (*d*)
24m Child Engagement	2.039	155	.043*	0.33	.442	158	.659	-0.070
36m Child Engagement	-1.716	130	.089	-0.29	-1.433	140	.154	0.233
24m BSID-II MDI	.192	152	.848	-0.03	.424	152	.672	-0.068
36m BSID-II MDI	-.634	146	.527	0.10	-.300	148	.765	0.049
24m CBCL	1.273	175	.205	-0.19	.967	178	.335	-0.144
36m CBCL	1.224	179	.223	-0.18	.128	178	.899	-0.019
36m PPVT	-.593	122	.555	0.11	-.329	128	.743	0.058

Note. Asterisk (*) indicates statistically significant difference at alpha = .10.

Table 21 displays the pilot results of the sensitivity analysis, which include an assessment of error type, assessment of overlap of the effect size confidence intervals, the magnitude of the difference between effect sizes found between threatened and randomized, the percent bias computed from the means, and an overall assessment of bias made from the preponderance of evidence from the preceding sensitivity analyses. The overall assessment of bias was rated “yes” for biased when the results for an individual outcome variable differed in the interpretation of the results from threatened to randomized.

Table 21

Foundation Small Condition 1 Pilot: Sensitivity Analyses for Point-in-Time Child Outcomes (Threatened vs. Randomized)

	Error Type	Overlapping Effect Size C.I.s	Effect Size Difference	Percent Bias	Overall Bias
	Type I or Type II	Y=yes, N=no	|d|>.20	%>10%	Y=yes, N=no
24m Child Engagement	Type I	Y	-0.40	0.43%	Y
36m Child Engagement	None	Y	0.52	-0.36%	Y
24m BSID-II MDI	None	Y	-0.04	-0.69%	N
36m BSID-II MDI	None	Y	-0.06	-0.74%	N
24m CBCL	None	Y	0.05	0.22%	N
36m CBCL	None	Y	0.16	-0.31%	N
36m PPVT	None	Y	-0.05	0.23%	N

Table 21 shows a Type I error, (i.e., a false statistically significant difference between intervention and control) for 24-month child engagement. In this case, for the threatened data, the control group mean is falsely shown to be higher than that of the intervention group (4.51 and 4.13, respectively). The effect sizes for the threatened data also show meaningful differences from randomized on 24 and 36 month child engagement. These larger effects changed the overall interpretation compared with the randomized data. Errors in the significance testing or meaningful effect size differences were not found for the other point-in-time measures. Percent bias was under 10% for these tests as well, indicating that overall threatened mean values did not differ from the randomized means. Another method of looking at the effect size differences was the computation of effect size confidence intervals. In this case, the confidence intervals showed the effect sizes overlapped (were not different), which conflicted with the more practical calculation of effect size differences which, for the two variables, showed a change in the interpretation from the randomized data.

Table 22 shows statistically significant differences between 24 and 36 months on the within groups child engagement for both threatened and randomized conditions. Thus, the sensitivity analysis shows no Type I or Type II errors. The situation is similar for the child behavior checklist means from 24 to 26 months. A notable difference was found for the effect size for the between groups (intervention versus control) on the CBCL. Table 23 shows that the effect size found for the threatened condition was highly inflated (by at least a medium effect; thus a very different interpretation) when compared with what was found for the randomized set.

Table 22

Foundation Small Condition 1: Repeated Measures Analysis of Variance Child Outcomes (Threatened vs. Randomized)

	Threatened Dataset 1			Randomized Dataset		
	Test Value	Sig Level (α=.05)	Effect Size (Cohen's d)	Test Value	Sig Level (α=.05)	Effect Size (Cohen's d)
Child Engagement (within)	15.663	<.001*	.684	17.164	<.001*	.708
Child Engagement (between)	.011	.918	.018	.813	.369	.154
BSID-II (within)	4.222	.042*	.362	3.251	.074	.316
BSID (between)	.410	.523	.113	.036	.851	.033
CBCL (within)	11.556	.001*	.523	9.835	.002*	.480
CBCL (between)	2.402	.123	.749	.448	.504	.102

Note. Asterisk (*) indicates statistically significant difference at alpha = .10.

Table 23

Foundation Small Condition 1 Pilot: Sensitivity Analyses for Change-Over-Time Child Outcomes (Threatened vs. Randomized)

	Error Type	Overlapping Effect Size C.I.s	ES Difference	Percent Bias	Overall Biased
	Type I or Type II	Y=yes, N=no	Criteria: |d| > .2	Criteria: % > .05	Y=yes, N=no
Child Engagement (within)	None	Y	0.024	0.85%	N
Child Engagement (between)	None	N	0.136	110.93%	Y
BSID-II (within)	Type I	Y	-0.046	-17.48%	Y
BSID (between)	None	Y	-0.080	-241.90%	Y
CBCL (within)	None	Y	-0.043	-11.60%	Y
CBCL (between)	None	N	-0.647	-146.87%	Y

Note. The confidence intervals are for the effect size estimates, not the mean estimates; thus, they are not indicative of the error type of the significance test which is based on the means.

Table 23’s summary of the sensitivity analysis is similar to what is described below in the results for the Phase III analysis. In the pilot, the threatened condition resulted in one Type I error and one highly inflated effect size. In addition, the percent bias showed over 10% bias (in some cases substantially over) for the threatened outcome data for five of the six repeated measures analyses. The negative sign on the percent bias means that the threatened data’s means are higher than the randomized data’s means.

The last set of pilot analyses (see Table 24) involved the use of propensity score weights to create adjusted estimates with respect to the randomized and threatened data. The adjustment did not correct the Type I error found for 24 month engagement but the effect size difference was over a small effect size (i.e., .20) less. For the 36 month child engagement, the estimate became worse and introduced a Type I error that was not found in the original threatened data, though again the effect size more closely resembled that of the randomized compared to the threatened data. Generally speaking, the propensity score-adjusted data were not helpful with regard to error rate but did improve effect size estimates in three cases and reduced percent bias in two cases. Table 24 is a preview of results found for Phase IV.

Table 24

Foundation Small Condition 1 Pilot: Effects Propensity Score Analysis and Analysis of Covariance Adjustments

	Bias Introduced in Threat Condition	Error Type	Assessment of Error Type	ES Difference	Assessment of ES Difference	Percent Bias Reduction	Percent Bias Assessment
	Y=yes, N=no	Type I or Type II	W=worse, B=better, NC=No change	[2]ESr-ESa	W=worse, B=better, NC=No change	pos=improved	Chg>10%
24m Child Engagement	Y	Type I	NC	-0.21	B	-0.46%	NC
36m Child Engagement	Y	Type I	W	-0.02	B	-2.06%	NC
24m BSID-II MDI	N	None	NC	0.07	NC	0.59%	NC
36m BSID-II MDI	N	None	NC	-0.02	NC	0.66%	NC
24m CBCL	N	None	NC	-0.06	NC	-1.32%	NC
36m CBCL	N	None	NC	-0.19	NC	-1.91%	NC
36m PPVT	N	None	NC	0.02	NC	-2.72%	NC
Child Engagement (within)	N	None	NC	-0.32	W	-20.00%	W
Child Engagement (between)	Y	None	NC	0.09	NC	-28.58%	W
BSID-II (within)	Y	Type I	NC	-0.22	W	-35.11%	W
BSID (between)	N	None	NC	-0.10	NC	-55.82%	W
CBCL (within)	Y	None	NC	-0.03	NC	1.67%	B
CBCL (between)	Y	Type I	W	-0.13	B	126.24%	B

A third and final purpose of the Phase I analysis (i.e., in addition to creating the foundation datasets and conducting the pilot) was to identify baseline characteristics that would be effective to use in the Phase IV corrective analyses. Each of the 12 primary baseline characteristics used in the sample balance and bias diagnostics were examined as well as any additional variables that were used for the exclusion processes. Multiple linear regression models were analyzed to determine the relationships among the baseline characteristics and outcome variables. Significant predictors included maternal risk, family resources, child gender, parent race, parent education, child age at enrollment, living arrangement, and child prematurity. Thus, these were initially selected for use in the corrective procedures.

Phase II: Threatened and Randomized Dataset Creation and Assessment

Creating Datasets

Recall the primary purpose of Phase II analyses was to address Research Question 1. The first step was to create each of the threatened and randomized conditions. Each of the datasets representing a threat condition was created using the foundation datasets created in Phase I using two different exclusion methods for each condition as specified in Chapter III. Unthreatened datasets, maintaining randomization, were also created from the corresponding foundation sets to use as comparisons by randomly selecting cases to match the threatened datasets on sample size (i.e., to eliminate sample size as a potential confound).

Six different variables were used in the exclusion processes, five of which were present in the original EHS data as individual case characteristics. The sixth variable, used in the noncompliance scenario, was created specifically for the current study using latent class analysis with variables related to EHS service compliance. Latent class models from one class to seven classes were consecutively estimated and compared. The model specification included variables related to the receipt of the following EHS services: case management, child care/development, parent support groups, regular weekly home visits, parenting classes, parent-child group activities, and group socialization events. Comparison of the models indicated a well-fitting five-class model of intervention group compliance in terms of multiple fit indices, parsimony, and interpretability. Results for the latent class analyses are shown in Tables 25 and 26.

Table 25

Compliance Groups Resulting from Latent Class Analysis

Class	Class Name	Proportion of Sample in Each Class	Probability of Class Membership
1	Very Low Service Compliance	.067	1.00
2	Highly Compliant Except Moderate on Home Visits	.204	.96
3	Highly Compliant Except Low on Home Visits	.137	.72
4	Moderately Compliant in All Services	.038	.76
5	Highly Compliant in All Services	.554	.87

Table 26

Model Fit Indices for Four-, Five-, and Six-Class Models of Latent Compliance

Fit Statistic	Four Classes	Five Classes	Six Classes
Log Likelihood	-2397.219	-2379.698	-2369.525
Akaike Info Criteria	4856.439	4837.395	4833.050
Bayesian Info Criteria (BIC)	5004.651	5023.857	5057.760
Sample Size Adjusted BIC	4906.202	4900.001	4908.498
Chi-square Test of Model Fit	192.601	101.53	49.602
Chi-square p-value	<.001	.137	.997
Likelihood Ratio (LR)	107.244	62.723	51.856
LR p-value	.203	.977	.994

The five-class model was selected on the basis of several indicators. First, the five-class model had a non-significant chi-square value compared to the significant chi-square of the four-class model; second, a more favorable adjusted BIC index was found for the five-class model compared to the six-class model; and third, the five-class model was interpretable while the six-class model was not interpretable and had two classes that consisted of 5% or less of the sample. As a final test, the resultant class variable was correlated with the second compliance exclusion variable (i.e., program-assessed program participation) to determine the degree of association between the two compliance-measuring variables (*r* = .498, *p* < .001). The resultant variable indicating latent compliant class was merged into the OBS prior to creating the three foundation datasets so that it would be present in all conditions for use in exclusion processes.

Assessing Datasets

Closely matching the procedures outlined in the original EHSRE study to test control-program group equivalence (thus assessing the degree of imbalance introduced in the threatened conditions), I performed univariate independent samples *t*-tests and chi-square tests to examine control versus program differences (using the 10% alpha level). In the case of these analyses, the criterion set for imbalance (i.e., differences between intervention and control participants in the threatened datasets) or bias (i.e., differences between participants in the threatened and randomized datasets) was if more than one baseline characteristic (of the 12 tested) was statistically significantly different.

In the Phase II section of this chapter, results are presented in the following order: (a) first, the combined results by manipulation are presented; then, (b) summary tables for the threat scenario results by individual condition are presented (i.e., for allocation bias, noncompliance, and differential attrition); and (c) summary tables indicating the level of balance retention for each of the randomized datasets matched on sample size are shown. Examples of the detailed tabular results for each individual condition tested in Phase II are presented in Appendix C because of space limitations. Appendix C includes examples of tables containing the means, standard deviations, frequencies for the baseline characteristics tested, *t*-test and chi-squares tests of significance comparing first intervention and control, and then threatened versus randomized values.

Phase II analyses were used to provide indication of whether *imbalance* in baseline characteristics was created within the threat conditions (i.e., statistically significant differences between *intervention and control* in proportions of the sample with selected characteristics) and across the scenarios. They also show whether *bias* was introduced within the threatened conditions (i.e., statistically significant differences found between the *threatened* and *randomized* sets in terms of the same set of characteristics) and across the scenarios.

Combined Summary Results
by Manipulation

Table 27 shows the overall number and proportion of tests that became imbalanced and biased with the introduction of the randomization threat. Representing the greatest proportion of tests out of balance, 100% of the nine method 1 allocation bias conditions resulted in intervention-control imbalance on the baseline characteristics. The scenario that had the lowest proportion of imbalanced conditions (44%) was differential attrition for method 1. Generally speaking, the proportions of conditions in which imbalance was introduced were greater than the proportion of bias between threatened and randomized that was found. Ultimately, this means that not all conditions resulted in imbalance and bias.

Table 27

Phase II Combined Summary Results for Baseline Imbalance and Bias by Scenario

Method	Threat Scenario	Intervention-Control Baseline Imbalance		Threatened-Randomized Baseline Bias	
		Avg number and rate (12) of sig diff tests per condition	Proportion of conditions in which imbalance was created (of 9)	Avg number and rate (of 12) of sig diff tests per condition	Proportion of conditions in which bias was introduced (of 9)
1	Allocation Bias	5.2 (43%)	100%	3.0 (25%)	67%
	Non-compliance	2.1(18%)	78%	2.0 (17%)	56%
	Differential Attrition	1.1 (9%)	44%	2.2 (19%)	33%
2	Allocation Bias	3.7 (31%)	78%	2.8 (23%)	67%
	Non-compliance	1.9 (16%)	56%	2.4 (20%)	56%
	Differential Attrition	3.0 (25%)	89%	2.8 (23%)	56%

Descriptively, the findings across the two methods for allocation bias (and similarly for the two methods for noncompliance) were fairly alike, indicating they were similar in terms of the magnitude of imbalance and bias created in the samples. The two methods of exclusion for differential attrition, however, yielded different results in terms of overall intervention-control imbalance created (44% vs. 89% of conditions). The differential attrition conditions overall resulted in less imbalance and bias in baseline characteristics than the other two threat scenarios. However, even at the lowest level of imbalance and bias observed across scenarios and methods (i.e., the imbalance bias introduced in the differential attrition conditions using method 1), 44% of the conditions resulted in intervention versus control imbalance and 33% resulted in threatened versus randomized bias on 12 baseline characteristics.

Sample size mattered in terms of data sets' susceptibility to imbalance and bias on baseline characteristics, given the results shown in Table 28. The pattern showed that as sample sizes increased, the intervention-control imbalance and the threat bias increased; however, this was reversed for the small and medium sets created using method 2 (see Table 29). No striking method differences were observed. The average number of tests that differed between intervention and control group increased as sample size increased; similarly, the number of tests that differed between threatened and randomized data increased as sample size increased. This was somewhat unexpected in the sense that researchers assume that larger datasets might be more resistant to imposed problems.

Table 28

Phase II Combined Summary Results for Baseline Imbalance and Bias by Sample Size

Method	Sample Size	Intervention-Control Baseline Imbalance		Threatened-Randomized Baseline Bias	
		Avg number and rate (12) of sig diff tests per condition	Proportion of conditions in which imbalance was created (of 9)	Avg number and rate (of 12) of sig diff tests per condition	Proportion of conditions in which bias was introduced (of 9)
1	Small	1.9 (16%)	67%	1.7 (14%)	56%
	Medium	3.0 (25%)	78%	2.7 (22%)	44%
	Large	3.6 (30%)	89%	2.9 (24%)	56%
2	Small	1.9 (16%)	78%	1.9 (16%)	44%
	Medium	3.2 (27%)	67%	2.9 (24%)	67%
	Large	3.4 (29%)	89%	3.2 (27%)	67%

Table 29

Phase II Combined Summary Results for Baseline Imbalance and Bias Based on Proportion of Sample Affected by Threat Condition

Method	Proportion of Sample Affected by Threat	Intervention-Control Baseline Imbalance		Threatened-Randomized Baseline Bias	
		Avg number and rate (12) of sig diff tests per condition	Proportion of conditions in which imbalance was created (of 9)	Avg number and rate (of 12) of sig diff tests per condition	Proportion of conditions in which bias was introduced (of 9)
1	25%	4.2 (35%)	100%	4.6% (38%)	100%
	15%	2.8 (23%)	67%	2.4 (20%)	56%
	5%	1.4 (12%)	56%	.22 (2%)	0%
2	25%	4.0 (33%)	100%	4.4 (37%)	89%
	15%	2.9 (24%)	78%	3.3 (28%)	89%
	5%	1.3 (11%)	44%	.22 (2%)	2%

The proportion of the sample affected by the threat condition had the biggest impact in terms of increasing imbalance and bias compared to sample size and threat type, particularly when the threat affected a large proportion of the sample. For all samples threatened at the 25% level, an average of 4.2 baseline characteristics became imbalanced between intervention and control and an average of nearly 5 was biased from the matched, randomized set. For the conditions at 25% threat using method 1, all (100%) of the datasets became imbalanced and biased. The sets, however, appeared to tolerate a 5%-of-the-sample-threat with minimal bias levels observed (0-2%). As with the sample-size manipulations, few method differences were observed with the exception of the 15% conditions introducing a greater proportion of bias in method 2. Individual condition summary results are provided by scenario in the sections below.

Phase II Allocation Bias Conditions

Table 30 shows the summary for each of the allocation bias conditions. Again, the individual analyses that made up this summary resembled those shown in the pilot Phase II analyses. Examples are also found in Appendix C.

Table 30

Summary of Phase II Allocation Bias Threat Conditions (Method 1)

	Imbalance in Baseline Characteristics (Intervention vs. Control)		Bias Introduced by Threat Condition (Threatened vs. Randomized)	
Condition	Number (and %) of Tests with Sig Difference	Imbalance Created	Number (and %) of Tests with Sig Difference	Bias Created
FSC1	5 (42%)	Yes	3 (25%)	Yes
FSC2	2 (17%)	Yes	3 (25%)	Yes
FSC3	2 (17%)	Yes	0 (0%)	No
FMC4	7 (58%)	Yes	4 (33%)	Yes
FMC5	7 (58%)	Yes	5 (42%)	Yes
FMC6	2 (17%)	Yes	1 (8%)	No
FLC7	9 (75%)	Yes	5 (42%)	Yes
FLC8	8 (67%)	Yes	5 (42%)	Yes
FLC9	5 (42%)	Yes	1 (8%)	No

As shown in the overall summary Table 27, 100% of the method 1 allocation bias conditions became imbalanced with the introduction of the threat condition. Table 30 displays the number (out of 12 tests) that were statistically significantly different on intervention and control (second column) and the number that were different between threatened and randomized data sets (fourth column). Foundation large condition 7 (FLC7; large sample with 25% threat level) had the greatest number of differences in terms of both imbalance and bias. The least affected condition was FSC3, the small dataset that had a 5% threat imposed; while imbalance was created, bias was not.

Method 2 allocation bias conditions (see Table 31) yielded similar findings; however, a few differences included the lack of imbalance and corresponding lack of bias created in FSC3 and FMC6 (two of the conditions in which only 5% of the sample was affected by the threat). The numbers of significant tests also varied slightly from method 1 to method 2.

Table 31

Summary of Allocation Bias Threat Conditions (Method 2)

	Imbalance in Baseline Characteristics (Intervention vs. Control)		Bias Introduced by Threat Condition (Threatened vs. Randomized)	
Condition	Number (and %) of Tests with Sig Difference	Imbalance Created	Number (and %) of Tests with Sig Difference	Bias Created
FSC1	3 (25%)	Yes	2 (17%)	Yes
FSC2	3 (25%)	Yes	5 (42%)	Yes
FSC3	0 (0%)	No	0 (0%)	No
FMC4	7 (58%)	Yes	5 (42%)	Yes
FMC5	4 (33%)	Yes	2 (17%)	Yes
FMC6	1 (8%)	No	0 (0%)	No
FLC7	7 (58%)	Yes	7 (58%)	Yes
FLC8	6 (50%)	Yes	3 (25%)	Yes
FLC9	2 (17%)	Yes	1 (8%)	No

Table 32 affirms that the randomized samples created to match the threatened conditions on sample size retained their balance, thus making them appropriate comparison sets for the Phase III and Phase IV analyses. While still in balance between intervention and control, only two of the conditions had zero significant differences.

Table 32

Retention of Balance in Random Subsamples of Randomized Datasets for Comparison with Allocation Bias Samples

Condition	Number (and %) of Tests with Sig Difference	Imbalance Created
FSC1-R	1 (8%)	No
FSC2-R	1 (8%)	No
FSC3-R	0 (0%)	No
FMC4-R	1 (8%)	No
FMC5-R	0 (0%)	No
FMC6-R	1 (8%)	No
FLC7-R	1 (8%)	No
FLC8-R	1 (8%)	No
FLC9-R	1 (8%)	No

Phase II Noncompliance Conditions

Similar to the allocation bias conditions, the following tables present summary results for the noncompliance conditions. For method 1 (see Table 33), imbalance was created for all of the 25%- and 15%-affected conditions. As with the allocation bias conditions, imbalance was not created in two of the 5%-affected conditions and bias was not created for any of the 5% conditions.

Table 33

Summary of Noncompliance Threat Conditions (Method 1)

	Imbalance in Baseline Characteristics (Intervention vs. Control)		Bias Introduced by Threat Condition (Threatened vs. Randomized)	
Condition	Number (and %) of Tests with Sig Difference	Imbalance Created	Number (and %) of Tests with Sig Difference	Bias Created
FSC10	3 (25%)	Yes	3 (25%)	Yes
FSC11	3 (25%)	Yes	2 (17%)	Yes
FSC12	0 (0%)	No	0 (0%)	No
FMC13	4 (33%)	Yes	5 (42%)	Yes
FMC14	2 (17%)	Yes	1 (8.3%)	No
FMC15	2 (17%)	Yes	0 (0%)	No
FLC16	3 (25%)	Yes	2 (17%)	Yes
FLC17	2 (17%)	Yes	5 (42%)	Yes
FLC18	0 (0%)	No	0 (0%)	No

Table 33 also shows that the method 1 noncompliance condition that was the most disrupted by the threat was FMC13, the medium sized data affected at the 25% level. It had 33% imbalance and 42% bias. Bias was not introduced in FMC14, which was affected at the 15% level. While generally similar results were found for the method 2 conditions (see Table 34), bias was introduced for FMC14 and a greater proportion of imbalance and bias was found for FLC16 (42% for both).

Table 34

Summary of Noncompliance Threat Conditions (Method 2)

	Imbalance in Baseline Characteristics (Intervention vs. Control)		Bias Introduced by Threat Condition (Threatened vs. Randomized)	
Condition	Number (and %) of Tests with Sig Difference	Imbalance Created	Number (and %) of Tests with Sig Difference	Bias Created
FSC10	2 (17%)	Yes	4 (33%)	Yes
FSC11	2 (17%)	Yes	1 (8%)	No
FSC12	1 (8%)	No	0 (0%)	No
FMC13	3 (25%)	Yes	6 (50%)	Yes
FMC14	1 (8%)	No	3 (25%)	Yes
FMC15	0 (0%)	No	0 (0%)	No
FLC16	5 (42%)	Yes	5 (42%)	Yes
FLC17	3 (25%)	Yes	3 (25%)	Yes
FLC18	0 (0%)	No	0 (0%)	No

The randomized comparison datasets for the noncompliance conditions retained their balance as desired. Three of the conditions had zero significant baseline differences between intervention and control (see Table 35).

Table 35

Retention of Balance in Random Subsamples of Randomized Datasets for Comparison with Noncompliance Samples

Condition	Number (and %) of Tests with Sig Difference	Imbalance Created
FSC10-R	1 (8%)	No
FSC11-R	0 (0%)	No
FSC12-R	0 (0%)	No
FMC13-R	0 (0%)	No
FMC14-R	1 (8%)	No
FMC15-R	1 (8%)	No
FLC16-R	1 (17%)	No
FLC17-R	1 (8%)	No
FLC18-R	1 (8%)	No

Phase II Differential Attrition Conditions

As mentioned earlier, the differential attrition conditions resulted in relatively less imbalance and bias than the other two threat scenarios and this is reflected in Table 36. Only one of the conditions affected at the 5% level became imbalanced and that was for the large dataset (FLC27). Imbalance and bias were uniformly introduced for all of the conditions affected at the 25% level in the method 1 conditions (i.e., FSC19, FMC22, and FLC25).

Table 36

Summary of Differential Attrition Threat Conditions (Method 1)

	Imbalance in Baseline Characteristics (Intervention vs. Control)		Bias Introduced by Threat Condition (Threatened vs. Randomized)	
Condition	Number (and %) of Tests with Sig Difference	Imbalance Created	Number (and %) of Tests with Sig Difference	Bias Created
FSC19	2 (17%)	Yes	3 (25%)	Yes
FSC20	0 (0%)	No	1 (8.3%)	No
FSC21	0 (0%)	No	0 (0%)	No
FMC22	3 (25%)	Yes	8 (67%)	Yes
FMC23	0 (0%)	No	0 (0%)	No
FMC24	0 (0%)	No	0 (0%)	No
FLC25	2 (17%)	Yes	7 (58%)	Yes
FLC26	1 (8.3%)	No	1 (8.3%)	No
FLC27	2 (17%)	Yes	0 (0%)	No

Method 2 (see Table 37) had twice as many conditions that resulted in imbalance compared with method 1 (i.e., eight out of nine, compared to four), appearing more effective as a threat method. Threat condition FMC22 (medium sample, 25% threat level) across both methods yielded the greatest number of statistically significant differences between the randomized and threatened sets. The high level of bias found in method 1 FLC25 (seven significantly different tests) was reduced to three significantly different tests in method 2.

Table 37

Summary of Differential Threat Conditions (Method 2)

	Imbalance in Baseline Characteristics (Intervention vs. Control)		Bias Introduced by Threat Condition (Threatened vs. Randomized)	
Condition	Number and %) of Tests with Sig Difference	Imbalance Created	Number (and %) of Tests with Sig Difference	Bias Created
FSC19	2 (17%)	Yes	1 (8.3%)	No
FSC20	2 (17%)	Yes	4 (33.3%)	Yes
FSC21	2 (17%)	Yes	0 (0%)	No
FMC22	5 (42%)	Yes	7 (58%)	Yes
FMC23	4 (33.3%)	Yes	3 (25%)	Yes
FMC24	4 (33.3%)	Yes	0 (0%)	No
FLC25	5 (42%)	Yes	3 (25%)	Yes
FLC26	1 (8%)	No	6 (50%)	Yes
FLC27	2 (17%)	Yes	1 (8.3%)	No

The randomized comparison datasets for the differential attrition conditions retained their balance as desired (see Table 38). As with noncompliance randomized sets, three of the condition had zero significant baseline differences between intervention and control.

Table 38

Retention of Balance in Random Subsamples of Randomized Datasets for Comparison with Differential Attrition Samples

Condition	Number (and %) of Tests with Sig Difference	Imbalance Created
FSC19-R	1 (8%)	No
FSC20-R	0 (0%)	No
FSC21-R	0 (0%)	No
FMC22-R	0 (0%)	No
FMC23-R	1 (8%)	No
FMC24-R	1 (8%)	No
FLC25-R	1 (8%)	No
FLC26-R	1 (8%)	No
FLC27-R	1 (8%)	No

Phase III: Sensitivity Analyses

The analytic Phase III was intended to answer Research Question 2a and 2b regarding the extent to which randomization disruption affected study results overall. For this phase, outcome analyses were performed using each threatened and randomized dataset and corresponding analyses were compared. The outcome analyses included the following components: (a) descriptive and frequency analysis to examine means and standard deviations and from which to compute effect sizes (Cohen’s d), effect size confidence intervals, effect size differences, and percent bias estimates; (b) independent samples *t*-tests to compare intervention and control groups on point-in-time outcomes;

and (c) repeated-measures ANOVA to test the main effects of change over time and differences between intervention and control groups.

Examples of the detailed outcome and sensitivity results tables for each individual condition are presented in Appendix D. Phase III summary results for each of the sensitivity analyses are shown below by (a) threat scenario (i.e., allocation bias, noncompliance, differential attrition); (b) sample size (i.e., small, medium, large); (c) proportion of sample affected by the threat (i.e., 5%, 15%, 25%); (d) by analysis type (i.e., point-in-time, repeated measures within groups, repeated measures between groups); and (e) child outcome measure (i.e., child engagement of parent, cognitive development, behavior problems, and receptive vocabulary). The summary results presented in this chapter were created by aggregating the sensitivity analyses in SPSS.

The sensitivity analyses were performed for all conditions and outcome measures. There were [(27 method 1 conditions + 27 method 2 conditions) X 13 outcome analyses] for a total of 702 outcome analyses results for which the set of four types of sensitivity analyses were performed. The assessment of Type I and Type II error rates was intended to determine whether the findings of the intervention vs. control significance test in the threatened conditions would differ from randomized data. The assessment of effect size differences (i.e., between threatened and randomized, the magnitude of the effect size differences and the rate of non-overlap of the effect size confidence intervals) was conducted to see if there were practical differences in the magnitude of the effects. The mean percent bias assessment was intended to look at the magnitude of differences between threatened and randomized means. The overall bias assessment was made using the preponderance of evidence from the sensitivity analyses in terms of meeting bias criteria.

Table 39 below shows the array of sensitivity analysis by threat scenario for method 1 conditions. In alignment with the findings in Phase II, the higher rates of baseline characteristic imbalance and bias introduced in the allocation bias threat conditions appeared to have resulted in slightly higher rates of disruption of the outcome analytic results. The rates of Type I and Type II error were higher for allocation bias conditions and the overall rate of bias. All of the mean effect size differences were very small, indicating very little deviation from the randomized effects. The rates of percent bias were similar across the conditions and, overall, the disruption to the outcome results was low.

Table 39

Sensitivity Analyses by Threat Scenario (Method 1)

	Rate of Type I Error	Rate of Type II Error	Mean Effect Size Difference (SD)	Rate of Non-Overlap in ES Confidence Intervals	Mean Percent Bias (SD)	Overall Bias Rate
Allocation Bias	5.1%	6.8%	.06 (.09)	47.9%	7.6% (.48)	39.3%
Noncompliance	2.6%	6.0%	.02 (.02)	53.0%	7.6% (.34)	35.0%
Differential Attrition	2.6%	3.4%	.04 (.06)	29.1%	7.8% (.53)	28.2%

Findings across exclusion method 1 and 2 were relatively similar (see Table 40). No large differences across threat scenarios were found in terms of rates of Type I and Type II errors (all close to 5% level that one might find by chance). The rate of Type I errors was highest for noncompliance in the second exclusion method. Type II errors were greatest for allocation bias for both exclusion methods. Mean effect size differences were very small (not reaching Cohen's criteria for a small effect at .20) and equivalent across the scenarios. Likewise, mean percent bias was the same for all scenarios except for noncompliance scenarios created using exclusion method 2 that almost reached the 10% level criteria.

Table 40

Sensitivity Analyses by Threat Scenario (Method 2)

	Rate of Type I Error	Rate of Type II Error	Mean Effect Size Difference (SD)	Rate of Non-Overlap in ES Confidence Intervals	Mean Percent Bias (SD)	Overall Bias Rate
Allocation Bias	3.4%	6.8%	.06 (.08)	45.3%	7.98% (.47)	35.9%
Noncompliance	6.8%	3.4%	.05 (.03)	48.7%	9.60% (.65)	29.9%
Differential Attrition	1.7%	3.4%	.02 (.03)	26.5%	5.72% (.26)	24.8%

While the rate of non-overlapping effect size confidence intervals was lower for the differential attrition threat scenarios (e.g., around 30% compared to nearly 50% found in the other two scenarios), these differences in effects sizes overall did not meaningfully change the interpretation of the size of the effects. Given the uniformity of findings across threat scenarios, these results provided some evidence of the similarity of effects of use of the exclusion methods used to create these scenarios.

Examination of the results by sample size (see Table 41) indicated that the highest rate of Type I error was found for the small samples, while the highest rate of Type II error (12%) was found for the large samples. Mean effect size differences did not differ much across the sample sizes and were very small. The rate of non-overlap (also indicating effect size differences) in the confidence intervals increased with the size of the sample. The large sample size showed an "above-the-10%-criteria" level of percent bias. Again, this might contradict some researchers' expectations that large samples are more robust to threats. Equivalent across the three sample sizes, the rate of bias overall was about a third of the conditions.

Table 41

Sensitivity Analyses by Sample Size (Method 1)

	Rate of Type I Error	Rate of Type II Error	Mean Effect Size Difference (SD)	Rate of Non-Overlap in ES Confidence Intervals	Mean Percent Bias (SD)	Overall Bias Rate
Small	6.0%	1.7%	.05 (.09)	18.8%	1.46% (.55)	35.0%
Medium	.9%	2.6%	.03 (.05)	46.2%	9.44% (.38)	35.9%
Large	3.4%	12.0%	.03 (.05)	65.0%	12.04% (.42)	31.6%

Note: Foundation sample sizes were small = 250, medium = 600, and large = 1,400.

Sample size comparisons for the method 2 (see Table 42) show the higher rates of Type I and Type II errors for the large sample with the small sample Type I error slightly lower than for method 1. The rate of Type I error for the medium sample was very low for method 2 conditions. The overall bias rate was highest for the small sample at 35% taking all evidence into consideration. Another difference from the method 1 conditions was the lower mean percentage bias rate for the large sample (7.6% versus 12.0%).

Table 42

Sensitivity Analyses by Sample Size (Method 2)

	Rate of Type I Error	Rate of Type II Error	Mean Effect Size Difference (SD)	Rate of Non-Overlap in ES Confidence Intervals	Mean Percent Bias (SD)	Overall Bias Rate
Small	5.1%	2.6%	.05 (.08)	23.9%	4.87% (.50)	35.0%
Medium	.9%	2.6%	.04 (.07)	41.0%	10.8% (.51)	28.2%
Large	6.0%	8.5%	.04 (.05)	55.6%	7.6% (.45)	27.4%

Note: Foundation sample sizes were small = 250, medium = 600, and large = 1,400.

Table 43 shows the sensitivity analysis by threat level (method 1). Because of the relatively higher proportion of baseline imbalance and bias created in the conditions threatened at 25% of the sample, the equivalence of the rates of Type I error across threat proportions was somewhat unexpected. However, the increase in the rate of Type II error as the threat proportion increased was not surprising as was the larger mean percent bias rate. All of the mean effect size differences were very small with a slightly higher difference for the 25% group. Overall, the larger the proportion of the sample affected by the threat condition, the greater outcome results disruption. In this case, over 46% of the outcome analyses resulted in bias for the 25% group (over double the rate for the 5%-affected samples).

Table 43

Sensitivity Analyses by Proportion of Sample Affected by Threat (Method 1)

	Rate of Type I Error	Rate of Type II Error	Mean Effect Size Difference (SD)	Rate of Non-Overlap in ES Confidence Intervals	Mean Percent Bias (SD)	Overall Bias Rate
Small (5%)	3.4%	1.7%	.01 (.02)	23.1%	2.63% (.16)	22.2%
Medium (15%)	3.4%	6.8%	.03 (.04)	52.2%	5.99% (.33)	34.2%
Large (25%)	3.4%	7.7%	.07 (.10)	54.7%	14.32% (.70)	46.2%

The method 2 conditions (see Table 44) resulted in very similar findings for the proportion of sample affected by the threat. The rate of Type I and Type II errors was the same for the 25% threat group at 6.0% and the effect size differences were slightly higher for the 25% group. The mean percent bias rate was much higher for the 25% group than for the 15% and 5% groups. At 16.6%, this exceeded the criteria for overall bias; however, the composite judgment of overall bias was lower for the 25% group for method 2 compared with method 1.

Table 44

Sensitivity Analyses by Proportion of Sample Affected by Threat (Method 2)

	Rate of Type I Error	Rate of Type II Error	Mean Effect Size Difference (SD)	Rate of Non-Overlap in ES Confidence Intervals	Mean Percent Bias (SD)	Overall Bias Rate
Small (5%)	2.6%	1.7%	.01 (.02)	21.4%	2.68% (.14)	21.4%
Medium (15%)	3.4%	6.0%	.05 (.06)	52.1%	4.01% (.47)	31.6%
Large (25%)	6.0%	6.0%	.06 (.09)	47.9%	16.62% (.68)	37.6%

While analyzing data, I noticed that a differential pattern was emerging for the outcome analysis types in terms of the rates of bias introduced. Examining the summary of results in Table 45, it is clear that the means negatively influenced by the threat conditions were strongly biased by the between groups results in terms of the mean percent bias. Over 33% of the outcomes resulted in biased findings. This led to an overall bias rate of 81.5% for the findings that resulted from the between group repeated measures analysis, indicating exaggerated positive or negative findings in terms of intervention effectiveness.

Table 45

Sensitivity Analyses by Type of Outcome Analysis (Method 1)

	Rate of Type I Error	Rate of Type II Error	Mean Effect Size Difference (SD)	Rate of Non-Overlap in ES Confidence Intervals	Mean Percent Bias (SD)	Overall Bias Rate
Point-in-Time (*t*-tests)	3.2%	5.8%	.06 (.07)	45.0%	.09% (.02)	14.8%
Repeated Measures ANOVA (within)	3.7%	2.5%	.02 (.03)	42.0%	.33% (.22)	32.1%
Repeated Measures ANOVA (between)	3.7%	7.4%	.02 (.07)	40.7%	33.25% (.88)	81.5%

Somewhat similar findings emerged from the method 2 analyses (see Table 46). The point-in-time analyses yielded higher Type I and Type II error rates than the repeated measures within-groups tests but the between groups ANOVAs had the highest error rates. Again, a high mean percent bias rate indicated a high overall rate of bias for the results of the between-groups tests.

Table 46

Sensitivity Analyses by Type of Outcome Analysis (Method 2)

	Rate of Type I Error	Rate of Type II Error	Mean Effect Size Difference (SD)	Rate of Non-Overlap in ES Confidence Intervals	Mean Percent Bias (SD)	Overall Bias Rate
Point-in-Time (*t*-tests)	4.2%	4.2%	.07 (.08)	40.2%	.30% (.01)	12.2%
Repeated Measures ANOVA (within)	1.2%	3.7%	.01 (.02)	37.0%	2.11% (.13)	18.5%
Repeated Measures ANOVA (between)	6.2%	6.2%	.01 (.04)	43.2%	30.85% (.97)	84.0%

Although presented in Tables 47 and 48, the receptive vocabulary results should be interpreted with caution given that they represent a much smaller proportion of the outcomes examined (i.e., due to the fact that the PPVT results were only available at 36 months, it precluded the completion of 24 month and the repeated measures analysis). A slightly higher rate of Type I error was found for child engagement compared with the other outcome types and a much larger rate of Type II error was found compared to that of the BSID-II and CBCL results (although somewhat less of a difference in the method 2 data shown in Table 48). While initially the interest was in examining the effect of disruption on different developmental domains, this finding was likely more of an artifact of the data type (ordinal 7-point scale compared with the continuous data of the other measures). Regardless, this result was not borne out in the method 2 findings.

Table 47

Sensitivity Analyses by Type of Child Outcome (Method 1)

	Rate of Type I Error	Rate of Type II Error	Mean Effect Size Difference (SD)	Rate of Non-Overlap in ES Confidence Intervals	Mean Percent Bias (SD)	Overall Bias Rate
Cognitive Development	3.7%	1.9%	.03 (.05)	44.4%	8.19% (.48)	40.7%
Child Behavior	2.8%	1.9%	.03 (.07)	45.4%	6.83% (.46)	27.8%
Child Engagement of Parent	4.6%	11.1%	.04 (.08)	38.9%	9.77% (.49)	39.8%
Receptive Vocabulary*	0.0%	11.1%	.07 (.05)	48.1%	.24% (.01)	11.1%

*Only 7.7% of the analyses were of receptive vocabulary scores because these data were only available at the 36 month age point.

The method 2 results (see Table 48) show a larger mean percent bias for the BSID-II cognitive development scores for the threatened sets compared to the randomized data. Given the otherwise similar sensitivity findings across the outcome measures in method 1 and 2, and the issue with the PPVT data, no clear conclusions about differential effects of outcome type could be made.

Table 48

Sensitivity Analyses by Type of Child Outcome (Method 2)

	Rate of Type I Error	Rate of Type II Error	Mean Effect Size Difference (SD)	Rate of Non-Overlap in ES Confidence Intervals	Mean Percent Bias (SD)	Overall Bias Rate
Cognitive Development	2.8%	1.9%	.04 (.07)	42.6%	17.97% (.46)	36.1%
Child Behavior	2.8%	2.8%	.03 (.06)	33.3%	.54% (.52)	20.4%
Child Engagement of Parent	6.5%	7.4%	.04 (.07)	44.4%	6.63% (.52)	38.0%
Receptive Vocabulary	3.7%	11.1%	.08 (.08)	40.7%	.45% (.01)	14.8%

Phase IV: Corrective Procedures

The corrective procedures employed in the study included the commonly used method of using a covariate in the model as a statistical control to improve effect estimates. Another form of adjustment used in the study was the application of propensity score weights to improve intervention-control balance (tested via same baseline characteristics used in the study's Phase II) and subsequently to improve the accuracy of the statistical estimates. Adjusted estimates for outcome analyses for all three threat scenarios and their corresponding conditions were obtained using these two corrective methods.

The propensity score analysis was conducted as specified in Chapter III with the inclusion of the following covariates for the predictive logistic regression: family mobility, maternal risk, single parent, child gender, parent race, parent education level, child age at enrollment, family living arrangement, receipt of food stamps, teen mother, parent primary occupation, English language ability, child first born, family poverty, and receipt of Medicaid. The model fit, classification quality, intervention-control ratios of propensity scores and of squared residuals, and balance restoration were assessed each time the model was used with a condition. In addition, significant predictors were identified, although they varied across the conditions. Maternal risk, education level, and living arrangement were typically significant predictors in the model.

The model’s goodness of fit was assessed with the Hosmer and Lemeshow (2000) test and the quality of classification was set at a minimum criterion of 70% correctly classified. The intervention versus control ratios of propensity scores and that of the squared standardized residuals were close to 1 (demonstrating equivalence). In all of the imbalanced conditions, balance between intervention and control baseline covariates was restored.

This set of predictors worked consistently well for all conditions with a few exceptions. In those exception cases, variables were removed to improve the model. The exceptions included the following (variables removed to improve fit and performance of model are indicated in the parentheses): FLC7 method 1 (maternal risk) and method 2 (parent education level); and FSC10 method 1 (maternal risk) and 2 (living arrangement and poverty).

Originally, two other procedures were considered for the statistical adjustments. Complier average causal effects (CACE) analyses were proposed for one of the noncompliance adjustment techniques; however, it was discovered that the methods used to create the threatened conditions (i.e., the exclusion process) imposed a violation of a critical assumption of the analyses. The assumption specified that, due to the randomization processes, one could infer that the same proportions of participants by compliance type or pattern would be found in both the intervention and control groups. In other words, the analyses relied on the premise that the same proportion of participants who were compliers in the intervention group was likely to be found in the control group had they been offered the intervention. Unfortunately, the way the exclusion process manipulated compliers in the intervention group disrupted the natural proportion of compliers and noncompliers, which would not resemble the original proportions assumed in the control group. Because of this, I did not use this procedure.

Replacing CACE, I conducted the noncompliance outcome analyses with the propensity scores weights to correct for imbalance along with the originally proposed ANCOVA adjustment. In light of the fact that I also used those two procedures for the allocation bias adjustments, it seemed advantageous to create a balanced design by using the same two corrective procedures for all three threat scenarios. Thus, I proceeded with using PSA weights and ANCOVA for all conditions. I intended to additionally examine the adjusted results for differential attrition using multiple imputation; however, difficulties in running the multiple imputation pooling procedures following imputation with repeated measures ANOVA (attempted with three different software packages and consultation) prohibited completion of that analysis. Thus, the results shown here effectively enabled a balanced comparison of the two corrective procedures effects across all conditions.

The adjusted results are shown below for all threat conditions created using the method 1 exclusion process and were split by biased or unbiased data (i.e., where bias had been introduced, or not, by the threat as per Phase III sensitivity analyses). In this way, the effects when the corrective procedure was warranted could be viewed separately from the effects when the adjustment was superfluous. In addition, the findings were disaggregated by the same categories presented in the Phase III analyses, specifically by (a) threat scenario (i.e., allocation bias, noncompliance, differential attrition); (b) sample size (i.e., small, medium, large); (c) proportion of sample affected by the threat (i.e., 5%, 15%, 25%); (d) analysis type (i.e., point-in-time, repeated measures within groups, repeated measures between groups); and (e) by child outcome measure (i.e., child engagement of parent, cognitive development, behavior problems, and receptive vocabulary). Tabular results for the exclusion method 2 are presented within the chapter for the threat scenario comparisons only; in all other cases, any differences on the comparative results from exclusion method 2 are commented on in the paragraphs below each method 1 table.

The adjusted results, using the same criteria for determining bias in Phase III, were assessed in terms of whether the adjustment improved, worsened, or did not change the estimates for each of the point-in-time and the repeated measures analyses. Findings for both the propensity score-adjusted and the ANCOVA-adjusted analyses are presented side-by-side for comparison of their relative effects. Note that positive percent bias reduction is desired, meaning that the adjustment procedure reduced bias in the threatened means.

Effects of the Corrective/Adjustment Procedures

Table 49 shows the propensity score-adjusted (PSA) and the covariate-adjusted results for each of the three threat scenarios and corresponding conditions for which the outcome results were biased. Type I and Type II error rates reflect both uncorrected error and newly introduced error. The next two columns show the proportions of outcome analyses for which errors were corrected and the rate of new errors. For the PSA-adjusted results, the Type I error was comparatively higher for allocation bias threat conditions; however, no Type II error was observed. The rate of corrected errors superseded the rate of newly introduced errors, which resulted from overcorrection to the point of finding statistically significant results when they did not exist for the randomized sets. For the noncompliance (NC) and differential attrition (DA) conditions, the error rates remained high and about 14% of the outcome analyses resulted in new errors in the significance tests. Only 5.3% and 2.9%, respectively, of the NC and DA, PSA-adjusted results showed the correction of error.

A substantial number of the effect size differences became worse with the adjusted results; in particular, for DA, 45.7% of the ANCOVA-adjusted results yielded larger effect size differences (i.e., subtracted from the randomized effect sizes) compared to the threatened effect sizes. Propensity score-adjusted allocation bias conditions had a10% rate of effect size difference improvement, while the ANCOVA-adjusted allocation bias had a 14% improvement rate. The adjustment techniques were not effective in reducing effect size difference for noncompliance or in ANCOVA-adjusted DA conditions. Mean percent bias improved somewhat for the PSA-adjusted DA conditions and generally rates of improvement were higher than rates of worse mean bias.

Table 49

Adjusted Versus Threatened-Biased Sensitivity Analyses by Threat Scenario (Method 1)

Method	Threat Type	Type I Error Rate	Type II Error Rate	Change in Error Rate		Mean ES Diff (SD)	Change in ES Difference		Mean Percent Bias Reduction (SD)	Change in Percent Bias	
				B	W		B	W		B	W
PSA	AB	32.0%	0%	18.0%	22.0%	.09 (.10)	10.0%	8.0%	-4.2% (.10)	34.0%	10.0%
PSA	NC	15.8%	15.8%	5.3%	13.2%	.13 (.14)	2.6%	15.8%	-2.8% (1.46)	28.9%	26.3%
PSA	DA	22.9%	8.6%	2.9%	14.3%	.08 (.06)	8.6%	5.7%	7.8% (.41)	8.6%	5.7%
AC	AB	4.0%	20.0%	16.0%	12.0%	.12 (.13)	14.0%	14.0%	1.28 (.13)	18.0%	4.0%
AC	NC	5.3%	18.4%	7.9%	5.3%	.08 (.10)	0%	13.2%	-.41 (.03)	21.1%	21.1%
AC	DA	8.6%	11.4%	0%	0%	.05 (.08)	0%	45.7%	-1.90 (.11)	34.3%	25.7%

Note: PSA = propensity score analysis, weights adjustment, AC = ANCOVA adjusted; AB = allocation bias, NC = noncompliance, DA = differential attrition; and for or change columns B = better, W = worse and remaining percentage (of 100%) resulted in no change on the bias measure. Again, note that positive percent bias is desired, since it indicates a reduction in bias in the threatened means..

It is important to note that Table 50 shows the effects of statistical adjustment when the data were *not biased*; thus, these generally are undesired effects, potentially correcting results when they do not need adjustment. For PSA adjustment, no Type II errors were created for any of the three threat scenarios. This was not the case, however, for the ANCOVA adjusted results in which low rates of Type II error were found. Propensity score analysis tended to maintain a higher level of Type I errors than did ANCOVA, i.e., a higher rate of statistically significant differences between intervention and control groups was observed that was not found in the randomized data. For both corrective methods, a larger proportion of the error rates became worse rather than better, which is understandable given that these results were from the threatened sets that did not become biased by the threat. Thus, the correction created new bias.

Table 50

Adjusted Versus Threatened-Unbiased Sensitivity Analyses by Threat Scenario (Method 1)

Method	Threat Type	Type I Error Rate	Type II Error Rate	Change in Error Rate		Mean ES Diff (SD)	Change in ES Difference		Mean Percent Bias Reduction (SD)	Change in Percent Bias	
				B	W		B	W		B	W
PSA	AB	16.4%	0%	0%	16.4%	.06 (.10)	0.0%	3.0%	-4.18% (.10)	0.0%	14.9%
PSA	NC	12.7%	0%	0%	12.7%	.06 (.09)	0.0%	3.8%	-2.65% (.14)	0.0%	5.1%
PSA	DA	15.9%	0%	0%	15.9%	.06 (.07)	8.5%	24.4%	-3.20% (.17)	0.0%	8.5%
AC	AB	1.5%	6.0%	0%	7.5%	.12 (.13)	0.0%	20.9%	1.28% (.13)	3.0%	0.0%
AC	NC	3.8%	3.8%	0%	6.3%	.08 (.10)	0.0%	10.1%	-.41% (.03)	0.0%	2.5%
AC	DA	3.7%	8.5%	0%	11.0%	.05 (.08)	1.2%	4.9%	-1.90% (.11)	0.0%	3.7%

Note: PSA = propensity score analysis, weights adjustment, AC = ANCOVA adjusted; AB = allocation bias, NC = noncompliance, DA = differential attrition; and for or change columns B = better, W = worse and remaining percentage (of 100%) resulted in no change on the bias measure.

The most striking difference in the method 2 conditions (see Table 51) is the larger mean percent bias findings, both positive and negative, as seen in the PSA-adjusted allocation basis conditions (mean bias improved by over 33%) and the PSA-adjusted differential attrition conditions, in which mean bias worsened by 21.5%. These results and those in the other corrected conditions contributed to overall higher rates of improvement with regard to mean bias reduction; however, the proportion of analyses that became worse was also high across conditions (range from 6% to 37% worse). For the ANCOVA-adjustments, a higher rate of Type II errors was found in method 2.

Table 51

Adjusted Versus Threatened-Biased Sensitivity Analyses by Threat Scenario (Method 2)

Method	Threat Type	Type I Error Rate	Type II Error Rate	Change in Error Rate		Mean ES Diff (SD)	Change in ES Difference		Mean Percent Bias Reduction (SD)	Change in Percent Bias	
				B	W		B	W		B	W
PSA	AB	28.0%	0.0%	16.0%	24.0%	.08 (.08)	8.0%	4.0%	33.94% (2.0)	22.0%	20.0%
PSA	NC	23.7%	7.9%	2.6%	18.4%	.08 (.07)	8.6%	5.7%	7.03% (.56)	31.6%	23.7%
PSA	DA	20.0%	5.7%	5.7%	20.0%	.06 (.05)	2.6%	2.6%	-21.54% (.63)	11.4%	37.1%
AC	AB	4.0%	22.0%	4.0%	10.0%	.12 (.10)	6.0%	12.0%	2.34% (25)	10.0%	6.0%
AC	NC	5.3%	13.2%	2.6%	5.3%	.10 (.11)	0.0%	7.9%	-9.24% (.42)	18.4%	21.1%
AC	DA	2.9%	8.6%	2.9%	2.9%	.18 (.31)	0.0%	25.7%	-1.89% (.17)	14.3%	17.1%

Note: PSA = propensity score analysis, weights adjustment, AC = ANCOVA adjusted; AB = allocation bias, NC = noncompliance, DA = differential attrition; and for or change columns B = better, W = worse and remaining percentage (of 100%) resulted in no change on the bias measure.

The unbiased data shown in Table 52 were subjected to quite a bit of unnecessary adjustment, alternately making improvements and making results more biased for almost all conditions for both adjustment types. In terms of error rates, effect size differences, and percent bias, higher proportions of analyses resulted in more biased outcome estimates compared with those that were improved. Very little (unneeded) improvement was found for effect size differences, which in this case was a positive finding. The rate of increased (worse) effect size differences was high for two of the allocation bias and noncompliance PSA conditions and two of the ANCOVA conditions. Mean bias reduction was proportionally less for the unbiased analyses compared with the biased, indicating greater efficacy of the adjustment for the biased results.

Table 52

Adjusted Versus Threatened-Unbiased Sensitivity Analyses by Threat Scenario (Method 2)

Method	Threat Type	Type I Error Rate	Type II Error Rate	Change in Error Rate		Mean ES Diff (SD)	Change in ES Difference		Mean Percent Bias Reduction (SD)	Change in Percent Bias	
				B	W		B	W		B	W
PSA	AB	31.3%	0.0%	0%	31.3%	.09 (.14)	1.5%	13.4%	-7.34% (.36)	1.5%	13.4%
PSA	NC	24.1%	1.3%	0%	22.8%	.07 (.06)	7.3%	22.0%	-5.15% (.21)	0.0%	10.1%
PSA	DA	15.9%	0.0%	0%	14.6%	.05 (.05)	1.3%	2.5%	-.32% (.12)	1.2%	6.1%
AC	AB	4.5%	3.0%	0%	7.5%	.11 (.13)	0.0%	20.9%	-1.48% (.13)	3.0%	0.0%
AC	NC	2.5%	5.1%	0%	3.8%	.10 (.10)	2.5%	10.1%	.44% (.08)	1.3%	2.5%
AC	DA	6.1%	13.4%	0%	14.6%	.08 (.17)	0.0%	7.3%	-.82% (.06)	2.4%	2.4%

Note: PSA = propensity score analysis, weights adjustment, AC = ANCOVA adjusted; AB = allocation bias, NC = noncompliance, DA = differential attrition; and for or change columns B = better, W = worse and remaining percentage (of 100%) resulted in no change on the bias measure.

.

In Table 53, the findings are presented by sample size. Sample-size comparisons of adjustment effects for biased data showed greater vulnerability for small samples, both in terms of ability to improve and to create worse estimates. This was reflected in the high rate of introduction of error in the significance tests (22.7%) as well as in the high rates of improvement and worsening in terms of the changes in effect size differences. The PSA adjustment for the small sample size resulted in a large mean percent bias increase and corresponding rate of change for the worse for means.

Table 53

Adjusted Versus Threatened-Biased Sensitivity Analyses by Sample Size (Method 1)

Method	Sample Size	Type I Error Rate	Type II Error Rate	Change in Error Rate		Mean ES Diff (SD)	Change in ES Difference		Mean Percent Bias Reduction (SD)	Change in Percent Bias	
				B	W		B	W		B	W
PSA	L	21.6%	16.2%	18.9%	13.5%	.07 (.07)	5.4%	2.7%	9.78% (.22)	24.3%	2.7%
PSA	M	16.7%	2.4%	4.8%	14.3%	.07 (.06)	0.0%	4.8%	3.57% (2.67)	31.0%	26.2%
PSA	S	34.1%	4.5%	6.8%	22.7%	.15 (.14)	15.9%	20.5%	-17.14% (1.41)	31.8%	29.5%
AC	L	2.7%	29.7%	10.8%	0.0%	.25 (.45)	0.0%	24.3%	.24% (.45)	8.1%	13.5%
AC	M	2.4%	11.9%	0.0%	9.5%	.13 (.11)	0.0%	19.0%	-.93% (.26)	21.4%	11.9%
AC	S	11.4%	11.4%	15.9%	9.1%	.13 (.11)	15.9%	25.0%	11.02% (.49)	38.6%	20.5%

Note: PSA = propensity score analysis, weights adjustment, AC = ANCOVA adjusted; Foundation sample sizes were S = small (250), M = medium (600), and L = large (1,400); B = better, W = worse, NC = no change.

For the PSA-adjusted method 2 data, a higher number of worse error rates was found (compared to method 1 above) for the medium sample size (30.5% compared with 14.3%), while this rate was reversed for the small sample size (13.6% compared with 22.7%). The rate of increased bias in effect size differences was lower in the small sample for PSA-adjusted method 2 data (4.5% compared with 20.5% using method 1data). Methods 1 and 2 were comparable on change in mean percent bias for PSA-adjusted data.

Lower rates of effect size bias increase were found in method 2 for medium and small samples. With the ANCOVA-adjusted results, lower rates of effect size difference (improvements) were found for both medium and small sizes (2.3% and 9.1%, respectively, compared to the results above). A lower rate of percent bias improvement was found in the ANCOVA-adjusted method 2 findings (4.8 compared with 21.4% above).

Sample-size comparisons of adjustment effects with *unbiased* results seen in Table 54 show that the Type I error rate was highest for the PSA-adjusted medium samples. Notably, no Type II errors were found in the corrected data for the PSA group. The large samples adjusted with ANCOVA, however, did introduce Type II error to over 12% of outcome analyses. As with most of the findings in Phase IV, mean effect size differences were very small for all sample sizes, although bias was introduced in 6.7% (medium samples with ANCOVA adjustment) to 16.4% (small samples adjusted with PSA) of outcome analyses. Little differences were found in mean percent bias reduction; however, 15% of the PSA-adjusted small samples analyses introduced new mean bias.

Table 54

Adjusted Versus Threatened-Unbiased Sensitivity Analyses by Sample Size (Method 1)

Method	Sample Size	Type I Error Rate	Type II Error Rate	Change in Error Rate		Mean ES Diff (SD)	Change in ES Difference		Mean Percent Bias (SD)	Change in Percent Bias	
				B	W		B	W		B	W
PSA	L	5.0%	0.0%	0%	5.0%	.03 (.02)	1.3%	7.5%	-2.09% (.13)	0.0%	3.8%
PSA	M	24.0%	0.0%	0%	24.0%	.05 (.06)	4.0%	9.3%	-4.05 (.19)	0.0%	9.3%
PSA	S	16.4%	0.0%	0%	16.4%	.10 (.13)	4.1%	16.4%	-3.84 (.09)	0.0%	15.1%
AC	L	0.0%	12.5%	0%	10.0%	.07 (.09)	0.0%	13.8%	.01% (.02)	0.0%	0.0%
AC	M	1.3%	1.3%	0%	2.7%	.07 (.09)	0.0%	6.7%	-1.27% (.10)	0.0%	4.0%
AC	S	8.2%	4.1%	0%	11.0%	.09 (.13)	1.4%	13.7%	-.11% (.14)	2.7%	2.7%

Note: PSA = propensity score analysis, weights adjustment, AC = ANCOVA adjusted; Foundation sample sizes were S = small (250), M = medium (600), and L = large (1,400); B = better, W = worse, NC = no change.

A few differences were found in the method 2 unbiased findings. For the large samples adjusted with PSA, an increased rate of bias was found using method 2 data (13.8% compared to 5.0%). Also, the rate of Type I error for the PSA-adjusted small samples was much worse at 28.8% compared to 16.4%. With respect to mean percent bias reduction in ANCOVA method 2 data, small samples had much higher rates of better and worse change compared with method 1, i.e., 27.3% of adjustments improved the percent bias while 15.9% became worse within the method 2 data.

The proportion-of-sample-affected comparisons of adjustment effects for *biased* outcome data revealed high rates of Type I error in the PSA-adjusted conditions (see Table 55). While some Type I and Type II error was reduced from the threatened sample, a larger proportion of new error was introduced yielding biased estimates. For the ANCOVA-adjusted 25% group, the mean effect size difference reached the criteria for a small effect. In this case, this was not a positive finding as most of the effect size differences were larger than were found for the threatened data. While it was hoped to see a positive adjustment with these techniques, particularly given the large outcome-biasing effect of the 25%-threat conditions, any positive changes were overshadowed by the proportion of tests that became worse. On the other hand, for the 25%-threat conditions adjusted by PSA, the mean percent bias reduction was meaningful (over 10%), which was a positive result. In fact all the mean percent bias reduction values showed some positive bias reduction.

Table 55

Adjusted Versus Threatened-Biased Sensitivity Analyses by Proportion of Sample Affected (Method 1)

Method	Threat Proportion	Type I Error Rate	Type II Error Rate	Change in Error Rate		Mean ES Diff (SD)	Change in ES Difference		Mean Percent Bias (SD)	Change in Percent Bias	
				B	W		B	W		B	W
PSA	25%	23.5%	11.8%	7.8%	21.6%	.13 (.11)	9.8%	15.7%	12.16% (2.74)	29.4%	29.4%
PSA	15%	21.4%	7.1%	16.7%	11.9%	.09 (.09)	7.1%	4.8%	7.60% (.22)	33.3%	9.5%
PSA	5%	30.0%	0%	3.3%	16.7%	.06 (.09)	3.3%	6.7%	2.17% (.52)	23.3%	20.0%
AC	25%	5.9%	21.6%	5.9%	7.8%	.20 (.35)	5.9%	23.5%	1.31% (.32)	29.4%	17.6%
AC	15%	7.1%	19.0%	9.5%	2.4%	.17 (.22)	0%	21.4%	2.10% (.14)	23.8%	9.5%
AC	5%	3.3%	6.7%	13.3%	10.0%	.10 (.10)	13.3%	23.3%	10.00% (.50)	13.3%	20.0%

Note: PSA = propensity score analysis, weights adjustment, AC = ANCOVA adjusted; proportions of sample affected were Large (L) = 25%, Medium (M) = 15%, and Small (S) = 5%; B = better, W = worse, NC = no change.

.

Method 2 *biased* data analyses included the following differences from method 1 findings. For the 5% threat group adjusted with PSA, method 2 findings indicated a lower (i.e., better) rate of worse analyses in terms of error rate (6.7% compared with 16.7%), although method 1 findings indicated a higher improvement rate for the PSA 15% threat group. Proportions of error rate changes were generally lower in method 2 ANCOVA findings compared with method 1. For effect size differences, the PSA results were similar for methods 1 and 2; however, the method 2 ANCOVA 25%-threat group had a lower rate of worse effect size results than did method 1. For the 5%-threat group adjusted with PSA, the rate of worse results was much higher for method 2 at 46.7%. For ANCOVA-adjusted results, method 1 showed higher rates of improvement in percent mean bias than were found for method 2 comparisons.

With the *unbiased* data presented in Table 56, the comparison of effects by the proportion of the sample affected indicated no Type II errors in the PSA-corrected data. This means that Type II errors were not introduced; however, some moderate overcorrection occurred within the PSA-adjusted data in the form of the Type I errors to find significant differences when they did not exist in the randomized data. The ANCOVA procedures did not yield as high of Type I error rates, although they did have small rates of Type II error introduction. In terms of effect size differences, mean differences were equivalent across adjusted conditions. Unfortunately, the proportion of larger (worse) effect size differences was higher across the board for all conditions than the improvement rates. Not much effect was observed in terms of mean percent bias reduction (which was favorable since these were already unbiased) except in the case of the PSA-adjusted 15% and 25% groups’ moderately worse rates of mean percent bias (over 10%).

Table 56

Adjusted Versus Threatened-Unbiased Sensitivity Analyses by Proportion of Sample Affected (Method 1)

Method	Threat Proportion	Type I Error Rate	Type II Error Rate	Change in Error Rate		Mean ES Diff (SD)	Change in ES Difference		Mean Percent Bias (SD)	Change in Percent Bias	
				B	W		B	W		B	W
PSA	25%	10.6%	0%	0%	10.6%	.08 (.10)	4.5%	15.2%	-3.09% (.09)	0%	12.1%
PSA	15%	12.0%	0%	0%	12.0%	.06 (.07)	1.3%	12.0%	-6.00% (.22)	0%	10.7%
PSA	5%	20.7%	0%	0%	20.7%	.04 (.08)	3.4%	6.9%	-1.11% (.05)	0%	5.7%
AC	25%	3.0%	7.6%	0%	10.6%	.08 (.09)	1.5%	10.6%	1.50% (.12)	1.5%	0%
AC	15%	0%	5.3%	0%	5.3%	.08 (.12)	0%	10.7%	-1.84% (.12)	0%	5.3%
AC	5%	5.7%	5.7%	0%	8.0%	.08 (.10)	0%	12.6%	-.74% (.05)	1.1%	1.1%

Note: PSA = propensity score analysis, weights adjustment, AC = ANCOVA adjusted; proportions of sample affected were Large (L) = 25%, Medium (M) = 15% and Small (S) = 5%; B = better, W = worse, NC = no change.

Differences in findings from method 2 *unbiased* data included the following. A worse rate of error was found in method 2 for the PSA-adjusted 15%-threat group (24% compared with 12.0%). No mean percent bias differences in the PSA adjusted data were observed between methods 1 and 2. No differences in the rates of change observed for error rate, effect size differences, or mean percent bias were found for ANCOVA from method 1 to method 2.

Comparisons of adjustment effects for *biased* data by analysis type (see Table 57) revealed that the PSA adjustments were successful in mitigating a high proportion of error (27.6%) for the point-in-time analyses with minimal rates of estimates becoming worse. This was not the case for the repeated measure between-groups tests in which this finding was reversed (26.9% of tests had false significance results). For ANCOVA, the Type II error rates were comparatively higher than those for the PSA adjustments. The ANCOVA adjusted results were also inconsistent with regard to improved or worsened rates of change. A high improvement rate for the point-in-time analyses was found but for the repeated measures within-groups analyses, 25.9% of the analyses had imposed Type I and Type II errors.

A positive effect for improvements in effect size differences was observed for the point-in-time analyses. However, as with the error rate assessments, the results for the effects of adjustment on the repeated measures analyses were not favorable. The repeated measure between-groups analyses yielded both high rates of improvement and worsening for mean percent bias reduction. Overall, the PSA adjustment on the repeated measures between groups had a favorable effect on mean percent bias (19% improvement rate).

Table 57

Adjusted Versus Threatened-Biased Sensitivity Analyses by Analysis Type (Method 1)

Method	Analysis Type	Type I Error Rate	Type II Error Rate	Change in Error Rate		Mean ES Diff (SD)	Change in ES Difference		Mean Percent Bias (SD)	Change in Percent Bias	
				B	W		B	W		B	W
PSA	PIT	24.1%	13.8%	27.6%	3.4%	.09 (.06)	27.6%	3.4%	-.55% (.01)	3.4%	0%
PSA	RMW	14.8%	7.4%	0%	7.4%	.16 (.13)	0%	25.9%	-5.78 (.14)	18.5%	25.9%
PSA	RMB	28.4%	4.5%	6.0%	26.9%	.08 (.10)	1.5%	6.0%	19.07 (2.41)	44.8%	26.9%
AC	PIT	10.3%	27.6%	20.7%	0%	.09 (.08)	20.7%	0%	-.14 (.01)	0%	0%
AC	RMW	11.1%	25.9%	7.4%	25.9%	.19 (.12)	0%	48.1%	5.48% (.16)	40.7%	3.7%
AC	RMB	1.5%	9.0%	4.5%	1.5%	.19 (.34)	1.5%	22.4%	4.64 (.44)	26.9%	26.9%

Note: PSA = propensity score analysis, weights adjustment, AC = ANCOVA adjusted; PIT = point-in-time, RMW = repeated-measures ANOVA within, RMB = repeated-measures ANOVA between; B = better, W = worse.

Differences in findings from the method 2 *biased* data showed that the PSA-adjusted point-in-time analyses had 17.2% of results with increased rates of error in the significance test compared with method 1's 3.4%. The rate of worse effect size differences for the PSA-adjusted repeated-measures within-groups analyses was 14.8% in method 2 compared with 25.9% in method 1. The mean percent bias was lower in method 2 (7.4% compared to 18.5%) for the PSA-adjusted repeated-measures within analyses. In terms of the effects of the ANCOVA adjustments, a lower proportion of tests resulted in improvements of significance errors for the point-in-time analyses (6.9% compared with 20.7%). A lower rate of error introduction was found for repeated-measures within analyses corrected by ANCOVA in the method 2 data (14.8% compared with 25.9% in method 1). The rates of improvement and worsening seen across analysis types in method 1 in terms of effect size differences were less pronounced for method 2. In the ANCOVA repeated-measures within analyses, the method 2 rate of improvement in mean percent bias was substantially lower than that of method 1 (14.8% compared with 40.7%).

With *unbiased* data shown in Table 58, the PSA adjustments again were unnecessary; thus, any improvements or worse findings contributed to an increase in parameter estimate bias. Similar to what was found in Table 56 for the comparisons of the proportions of the sample affected by the threat, no Type II error was introduced in the PSA conditions although small Type I and Type II error rates were found in the ANCOVA conditions. Moderately high rates of Type I error were found for the PSA point-in-time and the repeated-measures between analyses. The proportion of analyses that resulted in new significance errors was highest for PSA compared with ANCOVA.

A mean effect size difference equivalent to a small effect (.20) resulted from the ANCOVA-adjusted repeated-measures within analyses, which was not a favorable finding as the change was in the wrong direction (effect sizes became more different from randomized than was found in the unadjusted, threatened condition). A large, unfavorable, negative mean percent bias was introduced in the repeated measures analyses adjusted with PSA (23.5% mean bias increase), which was also reflected in the rate of the increase in unfavorable mean percent bias estimates for this analysis type.

Table 58

Adjusted Versus Threatened-Unbiased Sensitivity Analyses by Analysis Type (Method 1)

Method	Analysis Type	Type I Error Rate	Type II Error Rate	Change in Error Rate		Mean ES Diff (SD)	Change in ES Difference		Mean Percent Bias (SD)	Change in Percent Bias	
				B	W		B	W		B	W
PSA	PIT	18.1%	0%	0%	18.1%	.05 (.05)	4.4%	11.9%	-.29% (.01)	0%	0.0%
PSA	RMW	3.7%	0%	0%	3.7%	.10 (.15)	0%	11.1%	-6.96% (.16)	0%	27.8%
PSA	RMB	21.4%	0%	0%	21.4%	.03 (.04)	0%	0%	-23.50% (.42)	0%	42.9%
AC	PIT	3.1%	5.0%	0%	7.5%	.04 (.03)	.6%	0%	-.05% (.004)	0%	0%
AC	RMW	3.7%	9.3%	0%	11.1%	.21 (.14)	0%	48.1%	-.44% (.07)	1.9%	1.9%
AC	RMB	0%	7.1%	0%	0%	.04 (.03)	0%	0%	-5.00% (.39)	7.1%	28.6%

Note: PSA = propensity score analysis, weights adjustment, AC = ANCOVA adjusted; PIT = point-in-time, RMW = repeated-measures ANOVA within, RMB = repeated-measures ANOVA between; B = better, W = worse.

Differences in findings from method 2 *unbiased* data indicated that the PSA point-in-time analyses' rate of increased error in significance results was 26.9% for method 2 compared with 18.1% in method 1. The assessment of the number of worse error rate for PSA-adjusted repeated-measured between analyses was 35.7% for method 2 compared with 21.4% for method 1. No method differences were found for changes in effect size differences or changes in mean percent bias reduction for the PSA or ANCOVA results.

The final set of analyses pertaining to the adjustment techniques was intended to compare the effectiveness of the adjustments by child outcome type. For the biased outcomes, the rates of bias after adjustment are presented in Table 59. In terms of PSA-adjusted estimates, the BSID-II data were the most negatively affected by the attempted correction. Over 22% of the analyses introduced significance test errors and the mean percent bias increase was nearly 17%. Propensity score analysis adjustment, however, had a favorable effect on the CBCL mean percent bias reduction (at 57.5%). Across the measures, PSA appeared to create or maintain a moderate amount of Type I errors, while these rates were less for the ANCOVA adjusted findings. The ANCOVA adjustment also had a positive impact on the mean percent bias reduction for the CBCL measure. However, the ANCOVA adjustment negatively increased the effect size differences for the BSID-II results. The improvements seen in the mean percent bias (for both PSA and ANCOVA results) were likely not worth the proportion of tests for which the estimates became more biased.

Table 59

Adjusted vs. Threatened-Biased Sensitivity Analyses by Child Outcome Measure (Method 1)

Method	Outcome Measure	Type I Error Rate	Type II Error Rate	Change in Error Rate		Mean ES Diff (SD)	Change in ES Difference		Mean Percent Bias (SD)	Change in Percent Bias	
				B	W		B	W		B	W
PSA	BSID	29.5%	4.5%	0%	22.7%	.09 (.10)	0%	11.4	-16.89% (1.29)	27.3%	22.7%
PSA	CBCL	26.7%	3.3%	0%	20.0%	.12 (.12)	3.3%	16.7	57.50% (3.17)	40.0%	33.3%
PSA	CE	17.8%	11.1%	22.2%	11.1%	.09 (.09)	17.8%	4.4%	2.84% (.54)	26.7%	11.1%
PSA	PPVT	25.0%	25.0%	50.0%	0%	.09 (.03)	0%	0%	-1.00% (.01)	0%	0%
AC	BSID	6.8%	9.1%	4.5%	6.8%	.17 (.29)	2.3%	27.3%	1.55% (.16)	25.0%	9.1%
AC	CBCL	6.7%	10.0%	6.7%	13.3%	.21 (.40)	3.3%	23.3%	11.90% (.59)	36.7%	26.7%
AC	CE	4.4%	24.4%	13.3%	2.2%	.14 (.11)	8.9%	20.0%	.67% (.20)	15.6%	15.6%
AC	PPVT	0%	75.0%	25.0%	0%	.12 (.08)	25.0%	0%	0% (.00)	0%	0%

Note: PSA = propensity score analysis, weights adjustment, AC = ANCOVA adjusted; BSID = Bayley Scales of Infant Development-II Scale Score, CBCL = Achenbach Child Behavior Checklist, CE = Child Engagement with Parent, PPVT = Peabody Picture Vocabulary Test. Caution interpreting PPVT results, only *n* = 4 biased tests (thus, 25% represents one test).

Differences in findings from method 2 *biased* data included a lower rate found in method 2 analyses for the BSID effect size differences for PSA (2.3% were worse compared to 11.4% in method 1). For the PSA-adjusted BSID results, there was a higher rate of worse mean percent bias (34.1% compared with 22.7%). For the PSA-adjusted child engagement, a lower rate of improvement in mean percent bias was found (16.7% compared to 40%). For the ANCOVA adjustments, 0% of results were better for child engagement in terms of the error rate (compared with 13.3%). For ANCOVA-adjusted BSID results, the rate of results that were worse was 15.9% for method 1. Thus, method differences were inconsistent in terms of the direction of the effects.

Outcome measure comparisons of the (unneeded) adjustment effects for *unbiased* data shown in Table 60 reveal that, again, the PSA adjustment did not create new Type II error; rather, it improved those issues. However, it was at the expense of unbiased significance testing as Type I errors were more common than for the ANCOVA-adjusted results. The proportion of analyses for which a significance error was observed was equally poor for BSID and for child engagement. No large mean effect size differences were found for either ANCOVA or PSA for any of the child outcome measures, which was a positive finding. However, the rate of worse results in terms of effect size difference was equally high for both PSA- and ANCOVA-adjusted child engagement. The adjustments did not have a large effect on mean percent bias reduction, though again, the PSA-adjusted child engagement data were susceptible to a greater mean bias increase compared with the other measures.

Table 60

Adjusted Versus Threatened-Unbiased Sensitivity Analyses by Child Outcome Measure (Method 1)

Method	Outcome Measure	Type I Error Rate	Type II Error Rate	Change in Error Rate		Mean ES Diff (SD)	Change in ES Difference		Mean Percent Bias (SD)	Change in Percent Bias	
				B	W		B	W		B	W
PSA	BSID	17.2%	0%	0%	17.2%	.05 (.07)	0%	15.6%	-1.83% (.08)	0%	4.7%
PSA	CBCL	11.5%	0%	0%	11.5%	.06 (.08)	1.3%	1.3%	-5.41% (.21)	0%	11.5%
PSA	CE	17.5%	0%	0%	17.5%	.07 (.11)	9.5%	22.2%	-3.25% (.10)	0%	14.3%
PSA	PPVT	13.0%	0%	0%	13.0%	.04 (.04)	0%	0%	-.30% (.01)	0%	0%
AC	BSID	4.7%	3.1%	0%	6.3%	.05 (.06)	0%	3.1%	-.22% (.14)	3.1%	3.1%
AC	CBCL	5.1%	9.0%	0%	11.5%	.08 (.09)	1.4%	7.7%	-.35% (.10)	0%	2.6%
AC	CE	0%	6.3%	0%	6.3%	.12 (.15)	0%	28.6%	-.95% (.03)	0%	1.6%
AC	PPVT	0%	4.3%	0%	4.3%	.04 (.03)	0%	0%	-.04% (.002)	0%	0%

Note: PSA = propensity score analysis, weights adjustment, AC = ANCOVA adjusted; BSID = Bayley Scales of Infant Development-II Scale Score, CBCL = Achenbach Child Behavior Checklist, CE = Child Engagement with Parent, PPVT = Peabody Picture Vocabulary Test. Caution interpreting PPVT results, only *n* = 23 unbiased tests (thus, 4.3% represents one test).

Differences in findings from method 2 *unbiased* data showed that the PSA-adjusted CBCL data had a more unfavorable rate of analyses that resulted in significance test errors (30.8% compared with 11.5% for method 1). No other method differences were found in terms of effect size differences or mean percent bias reduction for the PSA-adjusted results. None of the method 2 findings differed for the measures using ANCOVA adjustments.

Chapter Summary

The introduction of randomization disruption was successful; generally speaking, the introduction of imbalance and bias in baseline characteristics led to bias in the results under threat conditions. The allocation bias scenario was most affected by imbalance under the threat condition; however, a high level of imbalance was also introduced in the noncompliance scenarios and a moderate amount in the differential attrition conditions. Significant baseline differences between threatened and randomized data by scenario were not as prevalent and ranged from affecting 33% to 67% of the tested differences. Baseline imbalance and bias were greatest for the large samples and for the samples that were threatened at the 25% threat level. As was expected, the greater the proportion of the sample affected by the threat scenario, the greater the likelihood of baseline imbalance, bias, and, subsequently, biased results.

The threat scenario under which the outcomes results were most sensitive was allocation bias, matching the larger baseline imbalance and bias introduced. Examination by sample size indicated a relatively high rate of Type I error was found for the small samples, while the highest rates of Type I and Type II error were among the large samples. Overall bias was highest among the small samples with 35% of the tests biased either in terms of the significance test, the mean effect size difference, or mean percent bias. The samples that were affected by the largest proportion of threat were the most sensitive to the disruption, again matching the levels of introduced baseline imbalance and bias. For this group, the high mean percent bias (14.3%) was notably higher than the 15% and 5% threat levels.

Overall, the adjustment techniques introduced more bias than they appeared to correct. The inconsistency in findings warrants much caution with regard to the reliance on these techniques for salvaging biased effect estimates. The final section of this work (Chapter V) includes a comprehensive discussion of the findings, their implications, and the recommendations for the field.

CHAPTER V

DISCUSSION

It is clear that research studies including the well-respected randomized design are susceptible to any number of design threats which, depending on the circumstances, might bias effect estimates of interest. Increasing understanding of when such threats might meaningfully obscure accurate and precise study findings is essential, particularly when stakes are high, e.g., when the results are intended to promote best practices in child and family programming. Downs et al. (2010) categorized potential randomization problems as occurring in the design phase (i.e., choice of randomization method), the implementation phase (i.e., allocation process), and errors occurring during the study. This study focused on elucidating understanding of the latter two problem types.

Overall, the current study's goals--to introduce and examine rates of imbalance and bias in baseline characteristics, to test the sensitivity of study results to such problems, and to attempt statistical adjustment of introduced bias--were mostly successfully achieved. This study has contributed steps toward furthering understanding of the effects of design threats on randomized study data. While extracting the salient findings from Chapter IV, several key themes emerged that deserve emphasis.

This discussion is organized by presenting the key findings for each phase of the analytic results, summarizing how the results are connected, and addressing what these results might indicate to the field. In addition, this chapter provides a discussion of challenges and issues that arose during the research and analytic process, shares how they were addressed, and offers recommendations for other educational and methodological researchers. Finally, the chapter closes with recommendations for future studies of this nature.

Key Findings and Implications of Phase II Results

The rate of imbalance introduced between intervention and control groups across the manipulated conditions ranged from 44% of conditions imbalanced in the method 1 differential attrition data to 100% of the nine conditions imbalanced by the method 1 allocation bias exclusion procedures. It is difficult to address whether this variability existed due to the different exclusion methods in and of themselves or the different variables used for exclusion. The differences in imbalance rates seen for method 2 attempts do not offer much insight into this question. More experimentation with this sort of exclusion process would clearly be helpful. However, it is notable that allocation bias was very clearly a selection bias threat and had an obvious connection to the imbalance created. Methodologically speaking, the noncompliance and attrition problems were less clearly associated with selection issues. To clarify, the manipulated variables were baseline characteristics related to whether a participant was included in the intervention versus the control group in the allocation bias scenario; whereas the noncompliance scenario manipulation was with data about program participation and the differential attrition scenario used variables collected about participants' resources at the 26-month time point.

Baseline differences for threatened versus randomized data (bias) were relatively lower than those achieved between intervention and control (imbalance). The results indicated that a range of 33% to 67% of the conditions became biased. Since the allocation bias conditions were more heavily threatened, it was expected that a high level of outcome bias would be found in the sensitivity analyses compared to the other two threat scenarios; that was indeed the case. With the same rationale and because the method 1 differential attrition data contained fewer conditions in which imbalance and bias were introduced, it was expected that less outcome bias would be observed in Phase III.

The categorical variables used in the balance and bias checking process generally seemed to be more influenced by the threat conditions than the continuous variables. Despite this finding, it is important in this type of diagnostic assessment to include a variety of categorical and continuous variables because it is not clear how consistent this pattern would be in different studies with alternative variable sets and conditions. A future study topic could focus on identifying commonly used variables across studies that tend to be more susceptible to balance disruption. Because of the scope of this study, I included a rather parsimonious list of baseline covariates; however, a greater number might have resulted in greater variability in the levels of imbalance and bias across the threat conditions. The variables and procedures selected for the balance check are important in identifying bias and their relative efficacy might differ across studies. In this vein, Arceneaux et al. (2004) warned about using incorrect randomization check procedures such as conducting the significance tests at the wrong level of analysis (e.g., checking at the individual level when randomization is at the group level as in cluster randomized studies).

Key Findings and Implications of Phase III Results

Phase III sensitivity results did differ by threat scenario, sample size, proportion of sample affected by the threat, analysis type, and child outcome measures. In terms of the threat scenarios, as mentioned earlier, the allocation bias outcome results were more sensitive to the threat conditions with higher Type I and Type II error rates and a higher overall bias rate than noncompliance and differential attrition, potentially for the reasons mentioned in the discussion of Phase II in the previous section. This finding was only weakened slightly in the method 2 results. Addressing Hewitt et al.'s (2010) suggestion that it is unclear when attrition becomes a serious threat to trial validity, the results here showed that bias occurred in approximately 25% of the outcome analyses. According to Foster and Bickman (1996), differential attrition is one of the greatest threats to internal validity; this certainly seemed to play out in this study.

The combined results of the Phase II and Phase III analyses led me to recommend that researchers and evaluators carefully avoid randomization threats that are so closely connected with the selection process, such as allocation bias based on participant characteristics or status. For example, a less obvious way that selection bias could be introduced into a randomized design includes allocating participants to intervention or control, informing them of their group status, and then delaying the commencement of the intervention for a substantial period of time. The promise of the intervention (or lack of) might cause the intervention and control groups to diverge in meaningful ways, e.g., the control group might seek other services immediately. This aligns with Dunford's (1990) indications that hidden threats to randomization might surface during the implementation and operation of experiments, e.g., ethical, legal, or liability issues that are also related to the selection process.

Given the goal of randomization to create intervention and control groups that were as similar as possible, any selection bias introduced might lead to violation of the assumption of strongly ignorable treatment assignment (Emura et al., 2008; Pearl, 2010; Shadish & Steiner, 2010). If this occurs, then group outcome differences might be incorrectly attributed to the intervention rather than a pre-existing imbalance.

Sample size comparisons showed that the highest rate of Type I error was found for the small samples, while the highest rate of Type II error was found for the large samples. The large-sample results indicated a slightly greater rate of mean percent bias introduced. The increased rate of Type II error for the large samples was curious given the increased power due to sample size to detect findings. Ultimately, the large sample rate of Type II error was not particularly high but as was stated in Chapter IV, the important note here is that large samples are not immune to threats and might in fact have unique problems (such as a greater likelihood of biased means) compared with smaller samples. The larger samples might potentially have produced greater mean bias because of increased variability on the outcome scores that was introduced with the threat conditions. Also, the threat condition itself eliminated much of the sample with the exclusion characteristic. The explanatory power of the exclusion characteristics appeared to have affected the outcome means in a way that made significant intervention versus control differences difficult to detect.

The biasing power of the largest (25%) proportion-of-the-sample-affected-by-the-threat condition was expected. This was especially true for the allocation bias scenarios and for the larger sample sizes. The introduction of a proportionally larger threat did indeed create relatively higher rates of Type I and Type II errors and elevated mean percent bias. The message here is that randomized studies might be able to (and do) tolerate a small amount of disruption. This is a subtle contrast to Gueron’s (2001) statement, indicating it is an “all-or-nothing” process; once problems are introduced, the study cannot recover. Clearly, the amount of bias they are able to tolerate is related to the maintenance of the integrity of other components of the research, such as measurement issues, and other issues that affect internal validity. The recommendation here is to prevent large threats to a study’s design and closely monitor the research activities to ensure rigorous practices are maintained. The reliance on randomization to prevent all internal validity problems should be in direct proportion to the energy put into maintaining a quality design.

The differences in levels of biased outcome results across analysis types also have implications for the selection of outcome analyses. It was somewhat surprising that the repeated measures between-groups analyses yielded higher rates of bias than the other types. A more in-depth investigation of the ways different analyses might emphasize or de-emphasize study bias is clearly warranted in the field. Similarly, this study showed that the selection of particular measurement strategies and tools could influence the masking or realization of study bias. The child engagement measure, a socio-emotional assessment of the children’s engagement with parent, was a play session coded and scored by video. More variability in the sensitivity results was observed of this measure compared with the Bayley Scales of Infant Development (1993) with a continuous range of scale scores.

Another study implication for consideration is that, to some extent, intervention studies attempting to improve outcomes that historically have small effect sizes in the literature are faced with a greater challenge of potentially losing barely detectable results in the face of small biasing effects. This might happen, for instance, when the intervention is not very powerful, e.g., when it has limited efficacy within a particular population or geographic region or is difficult to implement.

Key Findings and Implications of Phase IV Results

The Phase IV results for the study were at least consistent in the mutual introduction of new bias and the correction of old bias (i.e., originating in the threat condition). The results implied that much caution should be used in applying post-hoc procedures and that the selection of the best adjustment techniques for the problem be made (see challenges listed in the next section). While propensity score analysis (PSA) adjustment resulted in some improvements, the method used in this study overcorrected in many cases. Analysis of covariance (ANCOVA) was slightly more consistent with relatively lower rates of Type I and Type II errors. It seemed to be a much more conservative approach to control for extraneous variance.

It is not a good idea to adjust data unless there is very clear evidence that significant bias exists to warrant the risk of overcorrection. As was expected, all of the adjustment with unbiased conditions revealed the substantial introduction of problems. In the adjustment analyses, very small mean effect size differences were found regardless of threat scenario, sample size, proportion of threat, analysis type, and child outcome measure.

Interestingly, in the original Early Head Start Research and Evaluation (EHSRE) study, implementation of adjustment procedures with the full national dataset indicated the outcome estimates were fairly stable, likely because not much correction was needed unlike in the threat conditions employed in the current study (U.S. Department of Education, Administration for Children and Families, 2002a). Overall, the adjustment procedures were inconsistent across the comparisons and resulted in as much (or more) new bias as correction.

Methodological Successes and Challenges

Using the 10% alpha criterion for significance testing around imbalance was a choice initially based on an attempt to match the EHSRE's methods so there was continuity with the original study and so that expectations for imbalance could be set. Later, the question of the selected alpha level arose. Whereas the 5% level might initially have been perceived as the more conservative cutoff for significance, on reflection, the 10% alpha level was considered advantageous and more prudent with respect to the goal of the analyses. In other words, it is helpful to identify even minimal imbalance as it is not clear how much intervention and control differences could affect results in every circumstance. Thus, I selected the 10% alpha to increase the sensitivity of the tests to detect potential imbalance.

In Phase I of the study, the identification of latent compliance types using latent class analysis in M*plus* proved to be a successful method in modeling underlying compliance type based on patterns of program participation in the study. For programs of this sort, foreknowledge of participant compliance types might help support the development of extra supports for families who characteristically resemble others who have demonstrated low compliance behaviors with a given intervention. Ongoing interventions might benefit from the use of such analytic strategies to help craft effective programs for families.

One of the methodological challenges encountered in the study was in the development and execution of the exclusion methods to obtain the threatened datasets. Originally, I anticipated that the process would be nearly identical across the three threat scenarios; however, this was not the case. The primary difference was because, in the allocation bias scenario, I threatened both the intervention and the control groups; whereas, I only threatened the intervention group (and randomly subsampled from the control group) in the other two scenarios. Thus, while the pilot study was valuable in terms of identifying the general methods for creating the threat conditions, more specific details were not understood until faced with the challenges. Subtle differences across the scenarios were found in the appropriate way to obtain the proportions of threatened samples in intervention and control groups, which resulted in some backtracking.

For example, one challenge experienced during the exclusion method for differential attrition was that I needed to derive the baseline imbalance and bias estimates with the appropriate cases excluded; thus, as in the two other scenarios, I deleted them. However, I belatedly recognized that I needed the data for the cases to remain in the dataset to use for the later adjustment procedures. In this case, I replaced the excluded case data and created files in which only the 36-month data were excluded (the intended missing data per the differential attrition threat). Since I created this exclusion method myself and there were no clear guidelines in the literature about a process such as this, it was a bit of trial-and-error to obtain the desired conditions.

As discussed in Chapter IV, the noncompliance exclusion method inadvertently violated one of the assumptions of the complier average causal effect (CACE) analysis, which I believed would have ultimately been a more effective method for reducing bias introduced in the noncompliance threat conditions since it was specifically intended for that purpose. More foreknowledge related to this methodological clash would have been helpful given the lack of precedence for the exclusion processes used.

Likewise, there is a strong preference for conducting multiple imputation with missing data due to differential attrition. It is unfortunate that the outcome analyses proposed for this study were not well-matched to the multiple imputation procedures. Undoubtedly, using a multi-level or linear mixed effects model or growth curve modeling for the repeated measures investigation would have been more in line with the more modern multiple imputation procedure.

Given the rate of error and bias introduced in the PSA-adjusted outcomes and the inconsistency in terms of the effects of the adjustments in general, I suspect the use of propensity score weights might have not been as helpful as propensity score matching would have been. The reason the weighting method was selected was to enable the retention of as many cases in the data as possible, particularly since a part of the overall investigation involved examination of sample size. It is common that cases are lost in propensity score matching (i.e., when there is no good match); thus the decision was made to use weighting. In hindsight, the matching procedure might have been more effective in achieving intervention-control group balance on baseline characteristics and, ultimately, more effective in reducing bias in outcome estimates. In the correction of randomized data, suitable matches would likely be available given that the initial allocation would be at least partly successful in creating group equivalence. Because of the inconsistency of the PSA findings, I would recommend using alternative PSA methods to researchers facing identified imbalance between their intervention and control groups.

A typical concern in research design is whether the intervention and independent variables selected for use in the study adequately explain all the aspects of variation in the outcome variables. One limitation of this study was that I had only the use of the covariates the original EHSRE study measured. In this study, I needed to select covariates for the imbalance testing, some exclusion processes, and to develop the propensity scores. I included the top 12 covariates the original study deemed influential for the baseline testing and selected covariates for the exclusion process I believed were the best measures of the exclusion variables for the realistic scenarios I invented. Because the EHSRE study gathered so many covariates, there were many variables from which to choose and include in the propensity score models as it was favorable to use a "kitchen sink" (i.e., non-parsimonious) approach (Shadish & Steiner, 2010). Recommendations for choosing covariates include using those that are highly connected with the selection process (Steiner et al., 2010).

Additional Implications for Researchers and Evaluators

The execution of real-life evaluation studies in education is complex, involving numerous stakeholders, high-stakes decisions, and often underfunded budgets. Creating and conducting studies sufficiently rigorous (in terms of the ability to make valid causal claims) and in harmony with other aspects of the program and political climate is challenging indeed. While the system-level challenges might be tricky even in the best of circumstances, they could be compounded by research and design difficulties. For example, sample size is sometimes difficult to control if there are limitations to how many participants can be included in an intervention or if eligibility criteria are stringent. If it is impossible to obtain a large enough sample, researchers can assume they might have difficulty identifying imbalance. If significance tests cannot detect baseline imbalance, any bias or randomization failure might also not be detectable, leading to a greater vulnerability to biased study results, subsequent interpretation, and action related to the results.

Researchers and evaluators may never know if their results are valid and not biased in some way. It is difficult to detect the influence of study flaws that might lead to biased results. On the other hand, there are some precautions and preventative actions researchers can take. They might compare their study results with those from other similar studies and programs reported in the literature or by conducting meta-analyses. Unexpected findings might indicate randomization or program implementation problems. They might keep close track and document the quantity and severity of problems encountered during randomization and program implementation. Researchers might choose to implement the randomization procedures themselves rather than rely on program staff. This information could give researchers a “trustworthiness” factor with which to view results. Close monitoring and adherence to program fidelity could help reduce other biasing effects, i.e., researchers could help prevent/avoid noncompliance and attrition issues by closely communicating with program staff. Certainly conducting baseline group equivalence checks is important as well as testing balance in the sample for which complete data were collected. Finally, researchers could commit to the use of high quality measurement tools and indicators to help improve the accuracy of the findings. Each of these actions taken together could help improve the rigor and quality of the study, thus helping to ensure more accurate results.

A larger issue related to the rationale behind the selection of any given design is worth mentioning. Certainly there are circumstances in which an experimental design is not only non-feasible but will not answer the research questions of interest. Furthermore, there are situations in which an experimental design might be contraindicated based on the likelihood of failure given the complexity of the intervention, the program and research staff attitudes, skills, understanding, and overall support of such a design. In some cases, gaining program buy-in by using a quasi-experimental design might improve the quality of the research because staff would support allocation procedures and adhere more firmly to program implementation guidelines. In cases when the benefits of using a randomized design have limited chance of materializing, the trade-offs inherent in the design might not be worth it (e.g., withholding intervention from a control group without the global benefit of improved causal claims of the intervention’s effectiveness).

With the increase in methodological innovations in terms of improving researchers’ ability to estimate causal effects, researchers and programs could be less apologetic for using well-executed, highly effective quasi-experiments. This is particularly true in cases when a randomized design is not possible. More flexibility in federally funded educational research is needed so the overall quantity *and* quality of research is achieved. More strict standards about what constitutes strong evidence of program efficacy are needed at the federal level as well as clearer documentation of the circumstances under which a program is found to be effective. For example, when programs are recommended as best practices, it would be helpful if the research evidence was consistently accompanied by detailed population and geographic information as well as program implementation guidelines and fidelity measures. In addition, more transparent research implementation practices and pitfalls would help future program staff and researchers understand the limitations of the originally designed research.

Recommendations for Future Study

Using an actual real world dataset brought an authenticity to the complexity of the study and its findings. It is recommended that these findings be cross-checked and compared with those issued from other studies using real data and perhaps from those of Monte Carlo simulations. Understanding the ways in which threats to randomization affect precisely controlled, simulated data would help in setting clearer expectations and hypotheses for what would be found in complex real data derived from human subjects. Testing more specific directional hypotheses would also help isolate explanations for inconsistent results found in real life studies.

Another recommendation would be to examine a broader range of statistical corrective procedures to specifically increase knowledge about best practices for different threat types and study designs. Studying the comparative effects of an array of adjustment procedures with variously biased data would help reveal whether particular threats were better addressed with specific types of corrective procedures. Since the correct choice of corrective procedure could be vital to finding accurate and valid results, clearer guidelines are needed for matching adjustment procedures to the type and magnitude of bias.

Implications for Policy and Practitioners

It is recommended that practitioners, program leaders, and policy-makers use caution when adopting programs or interventions even when they are based on a seemingly rigorous evidence base. Simply trusting a study on the basis of having used a randomized design might lead to the implementation of programs that have not been used in a given region, school type, age group, or population in which program staff are interested. It might also be helpful to look for clues in published articles about the quality of the research and randomization implementation and adherence to quality control procedures. Good practice also involves selecting interventions based on more than one large rigorous study (replication of findings regarding an intervention’s effectiveness is important). It is critical to implement programs with demonstrated efficacy that were tested under a variety of conditions, circumstances, and budgets. Implementing the best programs with the highest chance of promoting positive change in education is an imperative we can no longer ignore.

REFERENCES

Achenbach, T. M., & Rescorla, L. A. (2000). *Manual for the ASEBA preschool forms and profiles*. Burlington, VT: University of Vermont, Department of Psychiatry.

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association, 91*(434), 444-455.

Angrist, J. D., & Krueger, A. B. (2001). Instrumental variables and the search for identification: From supply and demand to natural experiments. *The Journal of Economic Perspectives, 15*(4), 69-85. doi:10.1257/jep.15.4.69

Arceneaux, K. G., Alan, S., & Green, D. P. (2004). *Monte Carlo simulation of the biases in misspecified randomization checks.* New Haven, CT: Yale University Institution for Social and Policy Studies.

Atchade, Y., & Wantcheckon, L. (2005). *Noncompliance bias correction with covariates in randomized experiments.* Retrieved from http://ebookbrowse.com/atchade-wantchekon-pdf-d219747772

Baker, S. G., & Kramer, B. S. (2008). Randomized trials for the real world: Making as few and as reasonable assumptions as possible. *Statistical Methods in Medical Research, 17*(3), 243-252. doi:10.1177/0962280207080640

Barnes, S. A., Lindborg, S. R., & Seaman, J. J. W. (2006). Multiple imputation techniques in small sample clinical trials. *Statistics in Medicine, 25*(2), 233-245. doi:10.1002/sim.2231

Bayley, N. (1993). *Bayley scales of infant development, Second edition: Manual.* New York: The Psychological Corporation, Harcourt Brace & Company.

Berger, V.W. (2005). *Selection bias and covariate imbalances in randomized clinical trials.* London, UK: Wiley.

Berger, V. W., & Weinstein, S. (2004). Ensuring the comparability of comparison groups: Is randomization enough? *Controlled Clinical Trials, 25*(5), 515-524. doi:10.1016/j.cct.2004.04.001

Bijwaard, G. E., & Ridder, G. (2005). Correcting for selective compliance in a re-employment bonus experiment. *Journal of Econometrics, 125*(1), 77-111. doi:10.1016/j.jeconom.2004.04.004

Boruch, R., DeMoya, D., & Snyder, B. (2001). The importance of randomized field trials in education and related areas. In F. Mosteller& R. Boruch. (Eds.) *Evidence matters* (pp. 15-49). Washington, DC: Brooking Institution Press.

Bruhn, M., & McKenzie, D. (2009). In pursuit of balance: Randomization in practice in development field experiments. *American Economic Journal: Applied Economics, 1*(4), 200-232. doi:10.2307/25760187

Burtless, G. (2001). Randomized field trials for policy evaluation: Why not in education? In F. Mosteller & R. Boruch. (Eds.) *Evidence matters* (pp. 179-197). Washington, DC: Brooking Institution Press.

Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys, 22*(1), 31-72. doi: 10.1111/j.1467-6419.2007.00527.x

Campbell, D. T. (1969). Reforms as experiments. *American Psychologist, 24*, 409-29.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental design for research.* Chicago, IL: Rand McNally & Company.

Chen, H., Geng, Z., Zhou, X.-H., Small, D. S., Cheng, J., & Vansteelandt, S. (2009). Identifiability and estimation of causal effects in randomized trials with noncompliance and completely nonignorable missing data [with discussions and rejoinder]. *Biometrics, 65*(3), 675-691. doi:10.1111/j.1541-0420.2008.01120.x

Clark, M.H. (2008, November). *Practical applications of propensity scores* [PowerPoint slides]. Paper presented at the annual conference of the American Evaluation Association, Denver, CO.

Clason, D., & Mundfrom, D. (2012). Adjusted means in analysis of covariance: Are they meaningful? *Multiple Linear Regression Viewpoints, 38,* 8-15.

Cochran, W. G. (1957). Analysis of covariance: Its nature and uses. *Biometrics, 13*, 261-281.

Cook, T. D. (2002). Randomized experiments in educational policy research: A critical examination of the reasons the educational evaluation community has offered for not doing them. *Educational Evaluation and Policy Analysis, 24*(3), 175-199. doi:10.3102/01623737024003175

Cook, T. D. (2003). Why have educational evaluators chosen not to do randomized experiments? *The Annals of the American Academy of Political and Social Science, 589*(1), 114-149. doi:10.1177/0002716203254764

Cook, T. D., & Campbell, D.T. (1979). *Quasi-experimentation: Design and analysis issues for field settings.* Chicago, IL: Rand McNally College Publishing Company.

Cook, T. D., & Shadish, W. R. (1994). Social experiments: Some developments over the past fifteen years. *Annual Review of Psychology, 45*(1), 545-580. doi:10.1146/annurev.ps.45.020194.002553

Davies, R. S., Williams, D., & Yanchar, S. (2008). The use of randomisation in educational research and evaluation: A critical analysis of underlying assumptions. *Evaluation and Research in Education, 21*(4), 303-317. doi:10.1080/09500790802307837

De Anda, D. (2007). Intervention research and program evaluation in the school setting: Issues and alternative research designs. *Children & Schools, 29*(2), 87-94.

Dehue, T. (2001). Establishing the experimenting society: The historical origin of social experimentation according to the randomized controlled design. *The American Journal of Psychology, 114*(2), 283-302.

Downs, M., Tucker, K., Christ-Schmidt, H., & Wittes, J. (2010). Some practical problems in implementing randomization. *Clinical Trials, 7*(3), 235-245. doi:10.1177/1740774510368300

Dunford, F. W. (1990). Random assignment: Practical considerations from field experiments. *Evaluation and Program Planning, 13*(2), 125-132. doi:10.1016/0149-7189(90)90040-4

Dunn, G., Maracy, M., & Tomenson, B. (2005). Estimating treatment effects from randomized clinical trials with noncompliance and loss to follow-up: The role of instrumental variable methods. *Statistical Methods in Medical Research, 14*(4), 369-395. doi:10.1191/0962280205sm403oa

Dunn, L. M., & Dunn, L. M. (1997). *Peabody picture vocabulary test* (3rd ed.). Circle Pines, MN: American Guidance Services.

Egger, M. J., Coleman, M. L., Ward, J. R., Reading, J. C., & Williams, H. J. (1985). Uses and abuses of analysis of covariance in clinical trials. *Controlled Clinical Trials, 6*, 12-24. doi:10.1016/0197-2456(85)90093-5

Eid, M., Langeheine, R., & Diener, E. (2003). Comparing typological structures across cultures by multigroup latent class analysis: A primer. *Journal of Cross-Cultural Psychology, 34*(2), 195-210. doi:10.1177/0022022102250427

Education Sciences Reform Act of 2002, H. R. 3801-18) (2002). Retrieved from http://www.gpo.gov/fdsys/pkg/BILLS-107hr3801enr/pdf/BILLS-107hr3801enr.pdf

Emura, T., Wang, J., & Katsuyama, H. (2008). *Assessing the assumption of strongly ignorable treatment assignment under assumed causal models*. Retrieved from http://www.math.s.chiba-u.ac.jp/report/files/08008.pdf

Esterling, K. M., Neblo, M. A., & Lazer, D. M. J. (2011). Estimating treatment effects in the presence of noncompliance and nonresponse: The generalized endogenous treatment model. *Political Analysis, 19*(2), 205-226. doi:10.1093/pan/mpr005

Falaye, F. V. (2009). Issues in mounting randomized experiments in education research and evaluation. *Global Journal of Education Research, 8*, 21-27.

Fisher, R. A. (1925). *Statistical methods for research workers*. Edinburgh: Oliver and Boyd.

Fisher, R.A. (1935). *The design of experiments.* Edinburgh: Oliver and Boyd.

Fitzmaurice, G. M. (2003). Methods for handling dropouts in longitudinal clinical trials. *Statistica Neerlandica, 57*(1), 75-99. doi:10.1111/1467-9574.00222

Flick, S. N. (1988). Managing attrition in clinical research. *Clinical Psychology Review, 8*(5), 499-515. doi:10.1016/0272-7358(88)90076-1

Forsythe, A. B. (1987). Validity and power of tests when groups have been balanced for prognostic factors. *Computational Statistics and Data Analysis, 5,* 193-200.

Foster, E. M., & Bickman, L. (1996). An evaluator's guide to detecting attrition problems. *Evaluation Review, 20*(6), 695-723. doi:10.1177/0193841x9602000603

Foster, E. M., Fang, G. Y., & Conduct Problems Prevention Research Group. (2004). Alternative methods for handling attrition: An illustration using data from the Fast Track evaluation. *Evaluation Review, 28*(5), 434-464. doi:10.1177/0193841x04264662

Frangakis, C. E., & Rubin, D. B. (1999). Addressing complications of intention-to-treat analysis in the combined presence of all-or-none treatment-noncompliance and subsequent missing outcomes. *Biometrika, 86*(2), 365-379. doi:10.1093/biomet/86.2.365

Gall, M. D., Gall, J. P., & Borg, W. R. (2003). *Educational research: An introduction* (7th ed.). Boston, MA: Pearson Education, Inc.

Garraza, L. G., Azur, M., Stephens, R. L., & Walrath, C. M. (2011). Gender differences in patterns of child risk across programmatic phases of the CMHI: A multiple group latent class analysis (LCA). *The Journal of Behavioral Health Services & Research, 38*(2), 265-277. doi:10.1007/s11414-010-9219-6

Gauthier, S. M., Bauer, C. R. Messinger, D. S., & Closius, J. M. (1999). The Bayley scales of infant development-II: Where to start? *Journal of Developmental & Behavioral Pediatrics, 20*, 75-79.

Greenland, S. (2000). An introduction to instrumental variables for epidemiologists. *International Journal of Epidemiology, 29*(4), 722-729. doi:10.1093/ije/29.4.722

Grimes, D. A., & Schulz, K. F. (2002). Bias and causal associations in observational research. *The Lancet, 359*(9302), 248-252. doi:10.1016/s0140-6736(02)07451-2

Gueron, J. M. (2001). The politics of random assignment: Implementing studies and affecting policy. In F. Mosteller & R. Boruch (Eds.), *Evidence matters* (pp. 15-49). Washington, DC: Brooking Institution Press.

Guo, S., & Fraser, M.W. (2010). *Propensity score analysis: Statistical methods and applications*. Thousand Oaks, CA: Sage Publications.

Hagenaars, J. A., & McCutcheon, A. L. (2002). *Applied latent class analysis.* New York: Cambridge University Press.

Hall, N. L. S. (2002). *R. A. Fisher and randomized experimental design*. Retrieved from http://du.summon.serialssolutions.com/link/0/eLvHCXMwQ7QySSwPjExMgT1bI7QjuJEKezdRBlk31xBnD11YoRmfkpMTb24GTEamxsBMy9f0MFbmywyXfXv_7BW2-PrZBQBmsiu7

Hall, N. S. (2007). R. A. Fisher and his advocacy of randomization. *Journal of the History of Biology, 40*(2), 295-325. doi:10.1007/s10739-006-9119-z

Hansen, J., & Bowers, B.B. (2008). Covariates balance in simple, stratified and clustered comparative studies. *Statistical Science, 23,* 219-236.

Heckman, J. J. (2005). The scientific model of causality. *Sociological Methodology, 35*(1), 1-97. doi:10.1111/j.0081-1750.2005.00164.x

Hewitt, C. E., Kumaravel, B., Dumville, J. C., & Torgerson, D. J. (2010). Assessing the impact of attrition in randomized controlled trials. *Journal of Clinical Epidemiology, 63*(11), 1264-1270. doi:10.1016/j.jclinepi.2010.01.010

Hewitt, C. E., Torgerson, D. J., & Miles, J. N. V. (2006). Is there another way to take account of noncompliance in randomized controlled trials? *CMAJ: Canadian Medical Association Journal, 175*(4), 347-347. doi:10.1503/cmaj.051625

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81*(396), 945-960.

Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression*. New York: Wiley.

Hostetler, K. (2005). What is "good" education research? *Educational Researcher, 34*(6), 16-21. doi:10.3102/0013189x034006016

Imai, K., Keele, L., Tingley, D., & Yamamoto, T. (2011). Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies. *American Political Science Review, 105*(4), 765. doi:10.1017/s0003055411000414

Institute of Education Sciences, U.S. Department of Education, (2008). *Rigor and relevance redux: Director's biennial report to congress.* Retrieved from http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=IES20096010

Institute of Education Sciences, National Board for Education Sciences. (2011). *2011 annual report (NBES 2011-6005).* Retrieved from http://ies.ed.gov/director/board/reports/20116005/pdf/NBES_20116005.pdf

Jin, H., & Rubin, D. B. (2009). Public schools versus private schools: Causal inference with partial compliance. *Journal of Educational and Behavioral Statistics, 34*(1), 24-45. doi:10.3102/1076998607307475

King, G., Neilson, R., Coberley, C., Pope, J. E., & Wells, A. (2011). Avoiding randomization failure in program evaluation, with application to the Medicare health support program. *Population Health Management, 14,* S11-S22. doi:10.1089/pop.2010.0074

Kubitskey, B. W., Vath, R. J., Johnson, H. J., Fishman, B. J., Konstantopoulos, S., & Park, G. J. (2012). Examining study attrition: Implications for experimental research on professional development. *Teaching and Teacher Education, 28*(3), 418-427. doi:10.1016/j.tate.2011.11.008

Levi, I. (1983). The wrong box. *The Journal of Philosophy, 80,* 534-542.

Little, R. J., An, H., Johanns, J., & Giordani, B. (2000). A comparison of subset selection and analysis of covariance for the adjustment of confounders. *Psychological Methods, 5*, 459-476. doi:10.1037/1082-989x.5.4.459

Little, R. J., & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York: Wiley & Sons.

Little, R. J., & Yau, L. H. Y. (1998). Statistical techniques for analyzing data from prevention trials: Treatment of no-shows using Rubin's causal model. *Psychological Methods, 3*(2), 147-159. doi:10.1037/1082-989x.3.2.147

Liu, X. S. (2011). Covariate imbalance and precision in measuring treatment effects. *Evaluation Review, 35*, 627-641. doi:10.1177/0193841x12439195

Manolov, R., Solanas, A., Bulte, I., & Onghena, P. (2010). Data-division-specific robustness and power of randomization tests for ABAB designs. *Journal of Experimental Education, 78*, 191-214. doi:10.1080/0022097090329282

Marcellus, L. (2004). Are we missing anything? Pursuing research on attrition. *The Canadian Journal of Nursing Research, 36*(3), 82-82.

Marczyk, G. R., DeMatteo, D., & Festinger, D. (2005). *Essentials of research design and methodology*. Hoboken, NJ: John Wiley & Sons.

Maxwell, J. A. (2004). Causal explanation, qualitative research, and scientific inquiry in education. *Educational Researcher, 33*(2), 3-11. doi:10.3102/0013189x033002003

McKnight, P. E., McKnight, K. M., Sidani, S., & Figueredo, A. J. (2007). *Missing data: A gentle introduction*. New York: Guilford Press.

Mitra, N., & Heitjan, D. F. (2007). Sensitivity of the hazard ratio to nonignorable treatment assignment in an observational study. *Statistics in Medicine, 26*(6), 1398-1414. doi:10.1002/sim.2606

Muthén, L. K., & Muthén, B. O. (1998-2011). *Mplus user's guide* (6th ed.). Los Angeles, CA: Muthén & Muthén.

Mutz, D., & Pemantle, R. (2011). *The perils of randomization checks in the analysis of experiments.* Retrieved from http://www.math.upenn.edu/~pemantle/papers/Preprints/perils.pdf

National Research Council. (2002). *Scientific research in education.* Washington, DC: The National Academies Press.

National Institute of Child Health and Human Development, Early Child Care Research Network. (1997). Poverty and patterns of child care. In G. J. Duncan & J. Brooks-Gunn (Eds.), *Consequences of growing up poor* (pp. 100-131). New York: Russell Sage Foundation.

National Institute of Child Health and Human Development, Early Child Care Research Network. (1999). Child care and mother-child interaction in the first three years of life. *Developmental Psychology, 35*, 1399-1413.

No Child Left Behind. (2002). *A desktop reference.* Retrieved from http://www2.ed.gov/admins/lead/account/nclbreference/reference.pdf

Northwestern University. (2011). *Workshops on quasi-experimental design and analysis in education, 2011.* Retrieved from http://www.ipr.northwestern.edu/workshops/past-workshops/quasi-experimental-design-and-analysis-in-education/2011/index.html

Ong-Dean, C., Hofstetter, C., & Strick, B. R. (2011). Challenges and dilemmas in implementing random assignment in educational research. *American Journal of Evaluation, 32*(1), 29-49. doi:10.1177/1098214010376532

Papineau, D. (1994). The virtues of randomization. *The British Journal for the Philosophy of Science, 45*(2), 437-450.

Pearl, J. (2010). The foundations of causal inference. *Sociological Methodology, 40*(1), 75-149. doi:10.1111/j.1467-9531.2010.01228.x

Popper, K. (1968). *Conjectures and refutations.* New York, NY: Torchbooks.

Raikes, H., Green, B. L., Atwater, J., Kisker, E., Constantine, J., & Chazan-Cohen, R. (2006). Involvement in Early Head Start home visiting services: Demographic predictors and relations to child and parent outcomes. *Early Childhood Research Quarterly, 21*(1), 2-24. doi:10.1016/j.ecresq.2006.01.006

Raykov, T. (2010). Analysis of covariance. *Corsini Encyclopedia of Psychology*, 1-2. doi:10.1002/9780470479216.corpsy0053

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*(1), 41-55.

Rubin, D. B. (1972). *Estimating causal effects of treatments in experimental and observational studies*. Princeton, NJ: Educational Testing Services.

Rubin, D. B. (1977). Formalizing subjective notions about the effect of nonrespondents in sample surveys. *Journal of the American Statistical Association*, *72*(359), 538-543.

Ryan, K. E. (2004). Serving public interests in educational accountability: Alternative approaches to democratic evaluation. *American Journal of Evaluation, 25*(4), 443-460. doi:10.1177/109821400402500403

Sangi-Haghpeykar, H., Meddaugh, H. M., Liu, H., & Grino, P. (2009). Attrition and retention in clinical trials by ethnic origin. *Contemporary Clinical Trials, 30*(6), 499-503. doi:10.1016/j.cct.2009.06.004

Sassler, S., & McNally, J. (2003). Cohabiting couples' economic circumstances and union transitions: A re-examination using multiple imputation techniques. *Social Science Research, 32*(4), 553-578. doi:10.1016/s0049-089x(03)00016-4

Schafer, J. L. (1999). Multiple imputation: A primer. *Statistical Methods in Medical Research, 8*(1), 3-15. doi:10.1177/096228029900800102

Schochet, P., Cook, T., Deke, J., Imbens, G., Lockwood, J. R., Porter, J., & Smith, J. (2010). *Standards for regression discontinuity designs.* Retrieved from http://ies.ed.gov/ncee/wwc/pdf/wwc_rd.pdf

Sekhon, J. (2008). *The Neyman–Rubin model of causal inference and estimation via matching methods.* Oxford: Oxford University Press.

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

Shadish, W. R., Hu, X., Glaser, R. R., Kownacki, R., & Wong, S. (1998). A method for exploring the effects of attrition in randomized experiments with dichotomous outcomes. *Psychological Methods, 3*(1), 3-22. doi:10.1037/1082-989x.3.1.3

Shadish, W. R., & Steiner, P. M. (2010). A primer on propensity score analysis. *Newborn and Infant Nursing Reviews, 10*(1), 19-26. doi:10.1053/j.nainr.2009.12.010

Slavin, R., & Smith, D. (2009). The relationship between sample sizes and effect sizes in systematic reviews in education. *Educational Evaluation and Policy Analysis, 31*(4), 500-506.

Steiner, P. M., Cook, T. D., Shadish, W. R., & Clark, M. H. (2010). The importance of covariate selection in controlling for selection bias in observational studies. *Psychological Methods, 15*(3), 250-267. doi:10.1037/a0018719

Taylor, M. I. (2009, April). *What is sensitivity analysis?* London, UK: Hayward Medical Communications.

Taylor, M. I., & Innocenti, M. (1993). Why covariance? A rationale for using analysis of covariance procedures in randomized studies. *Journal of Early Intervention, 17,* 455-466. doi:10.1177/105381519301700409

Taylor, L., & Zhou, X. H. (2009). Multiple imputation methods for treatment noncompliance and nonresponse in randomized clinical trials. *Biometrics, 65*(1), 88-95. doi:10.1111/j.1541-0420.2008.01023.x

Towne, L., & Hilton, M. (2003). *Implementing randomized field trials in education.* Washington, DC: National Academies Press.

U.S. Department of Education, Administration for Children and Families. (2002a). *Making a difference in the lives of infants and toddlers and their families: The impacts of Early Head Start. Volume I: Final technical report.* Retrieved from http://www.acf.hhs.gov/programs/ opre/ehs/ehs_resrch/

U.S. Department of Education, Administration for Children and Families. (2002b). *Making a difference in the lives of infants and toddlers and their families: The impacts of Early Head Start. Volume II: Final technical report appendices.* Retrieved from http://www.acf.hhs.gov/ programs/opre/ehs/ehs_resrch/

U.S. Department of Education, Administration for Children and Families. (2006). *Research to practice: Early Head Start benefits children and families.* Retrieved from http://www.acf.hhs.gov/programs/opre/ehs/ehs_resrch/index.html

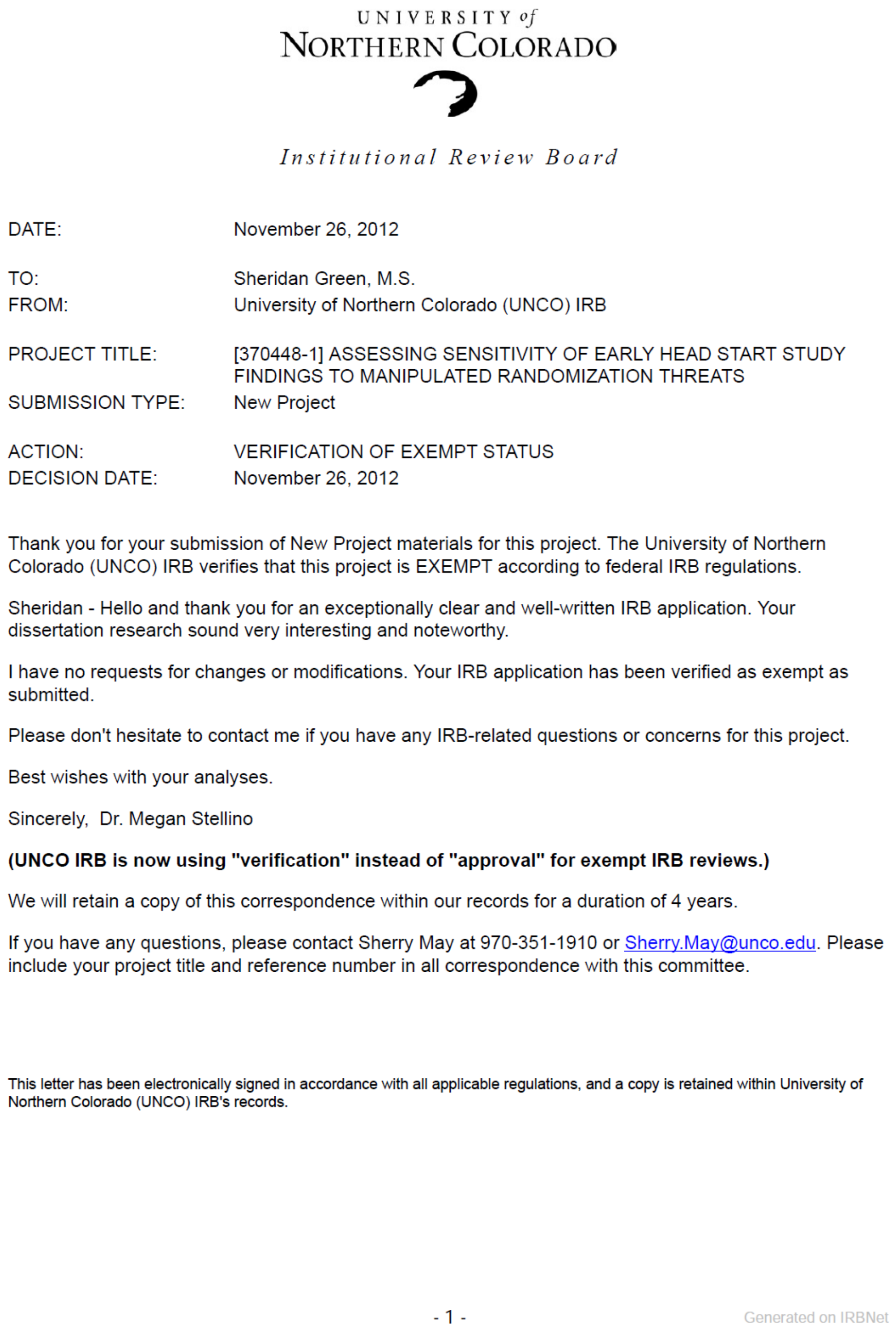
Van Breukelen, G. J. P. (2010). Analysis of covariance (ANCOVA). In N. J. Salkind (Ed.), *Encyclopedia of research design* (Vol. 1, pp. 20-26). Thousand Oaks, CA: SAGE Reference.

Vinovskis, M. (2002). Missing in practice? Development and evaluation at the U.S. Department of Education. In F. Mosteller & R. F. Boruch (Eds.), *Evidence matters: Randomized trials in education research* (pp.120-149). Washington, DC: Brookings Institution Press.

Yates, F. (1951). The influence of statistical methods for research workers on the development of the science of statistics. *Journal of the American Statistical Association*, *46*(253), 19-34.

APPENDIX A

INSTITUTIONAL REVIEW BOARD APPROVAL



UNIVERSITY of
NORTHERN COLORADO

Institutional Review Board

DATE:	November 26, 2012
TO:	Sheridan Green, M.S.
FROM:	University of Northern Colorado (UNCO) IRB
PROJECT TITLE:	[370448-1] ASSESSING SENSITIVITY OF EARLY HEAD START STUDY FINDINGS TO MANIPULATED RANDOMIZATION THREATS
SUBMISSION TYPE:	New Project
ACTION:	VERIFICATION OF EXEMPT STATUS
DECISION DATE:	November 26, 2012

Thank you for your submission of New Project materials for this project. The University of Northern Colorado (UNCO) IRB verifies that this project is EXEMPT according to federal IRB regulations.

Sheridan - Hello and thank you for an exceptionally clear and well-written IRB application. Your dissertation research sound very interesting and noteworthy.

I have no requests for changes or modifications. Your IRB application has been verified as exempt as submitted.

Please don't hesitate to contact me if you have any IRB-related questions or concerns for this project.

Best wishes with your analyses.

Sincerely, Dr. Megan Stellino

(UNCO IRB is now using "verification" instead of "approval" for exempt IRB reviews.)

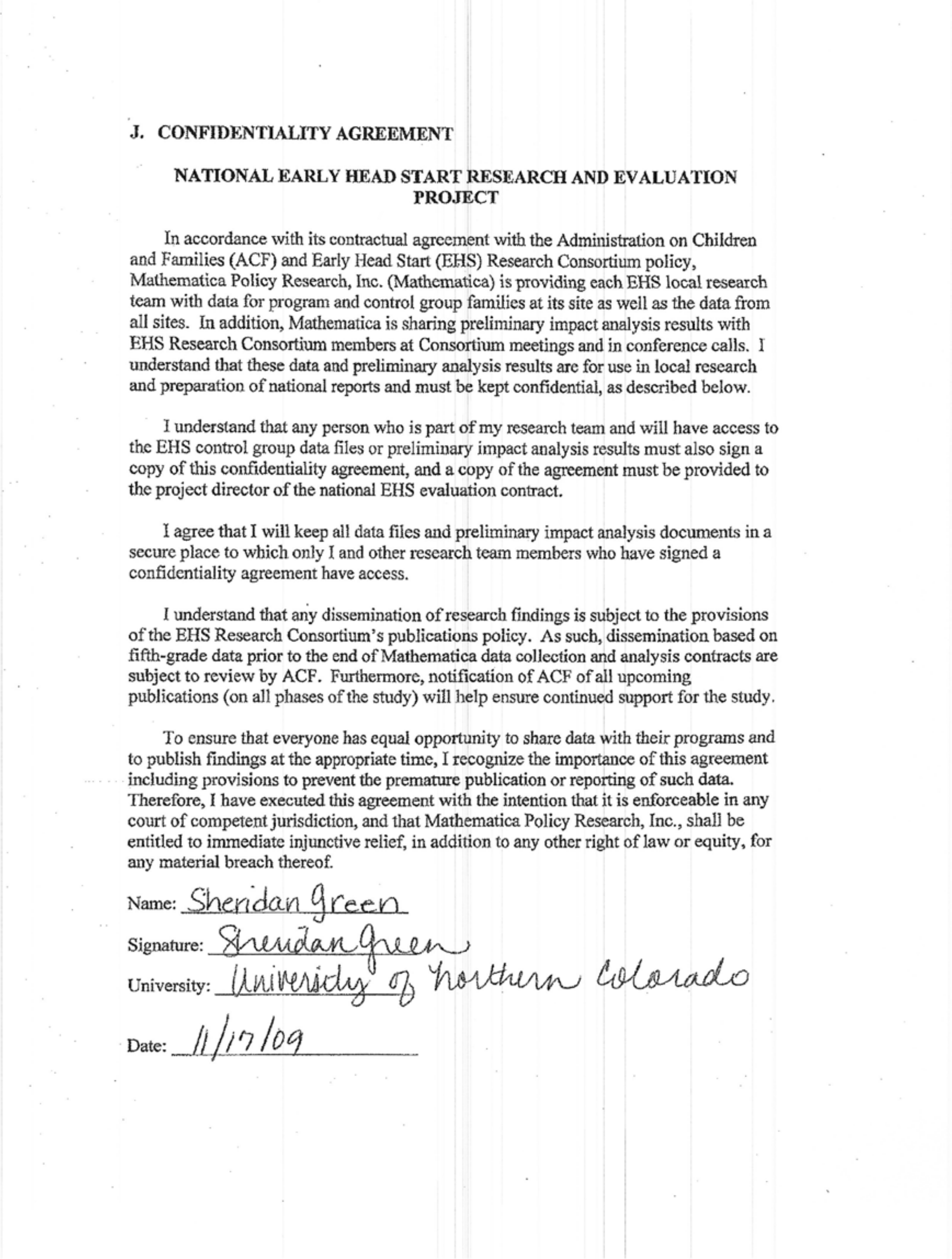
We will retain a copy of this correspondence within our records for a duration of 4 years.

If you have any questions, please contact Sherry May at 970-351-1910 or Sherry.May@unco.edu. Please include your project title and reference number in all correspondence with this committee.

This letter has been electronically signed in accordance with all applicable regulations, and a copy is retained within University of Northern Colorado (UNCO) IRB's records.

APPENDIX B

DATA USE AGREEMENT



J. CONFIDENTIALITY AGREEMENT

NATIONAL EARLY HEAD START RESEARCH AND EVALUATION PROJECT

In accordance with its contractual agreement with the Administration on Children and Families (ACF) and Early Head Start (EHS) Research Consortium policy, Mathematica Policy Research, Inc. (Mathematica) is providing each EHS local research team with data for program and control group families at its site as well as the data from all sites. In addition, Mathematica is sharing preliminary impact analysis results with EHS Research Consortium members at Consortium meetings and in conference calls. I understand that these data and preliminary analysis results are for use in local research and preparation of national reports and must be kept confidential, as described below.

I understand that any person who is part of my research team and will have access to the EHS control group data files or preliminary impact analysis results must also sign a copy of this confidentiality agreement, and a copy of the agreement must be provided to the project director of the national EHS evaluation contract.

I agree that I will keep all data files and preliminary impact analysis documents in a secure place to which only I and other research team members who have signed a confidentiality agreement have access.

I understand that any dissemination of research findings is subject to the provisions of the EHS Research Consortium's publications policy. As such, dissemination based on fifth-grade data prior to the end of Mathematica data collection and analysis contracts are subject to review by ACF. Furthermore, notification of ACF of all upcoming publications (on all phases of the study) will help ensure continued support for the study.

To ensure that everyone has equal opportunity to share data with their programs and to publish findings at the appropriate time, I recognize the importance of this agreement including provisions to prevent the premature publication or reporting of such data. Therefore, I have executed this agreement with the intention that it is enforceable in any court of competent jurisdiction, and that Mathematica Policy Research, Inc., shall be entitled to immediate injunctive relief, in addition to any other right of law or equity, for any material breach thereof.

Name: Sheridan Green

Signature: Sheridan Green

University: University of Northern Colorado

Date: 11/17/09

APPENDIX C

PHASE II DATA TABLES

Table 61

Means and Standard Deviations by Group for Foundation Small Condition 2 (Method 1), Foundation Small Condition 2 (Method 2), and Foundation Small Condition 2-Randomized on Five Continuous Baseline Characteristics

		FSC2			FSC2-2			FSC2-R		
	Group	n	Mean	SD	n	Mean	SD	n	Mean	SD
Mother Age	C	106	23.349	5.679	106	22.774	5.535	106	22.745	5.849
	I	105	22.248	6.295	105	22.276	6.257	105	22.590	6.437
Child Gender	C	106	.472	.502	106	.500	.502	106	.491	.502
	I	106	.575	.497	106	.528	.502	106	.585	.495
Child Premature	C	72	.125	.333	71	.113	.318	71	.099	.300
	I	75	.160	.369	77	.130	.338	77	.156	.365
Welfare Receipt	C	86	.244	.432	85	.306	.464	84	.298	.460
	I	80	.250	.436	81	.222	.418	84	.202	.404
Food Stamps Receipt	C	99	.364	.483	100	.450	.500	100	.430	.498
	I	101	.376	.487	101	.366	.484	103	.350	.479

Table 62

Frequencies by Group for Foundation Small Condition 2 (Method 1), Foundation Small Condition 2 (Method 2), and Foundation Small Condition 2-Randomized for Race

		FSC2					FSC2-2					FSC2-R				
Race	Group	A	B	C	O	Total	A	B	C	O	Total	A	B	C	O	Total
	C	43	26	29	6	104	46	32	22	4	104	44	26	28	6	104
	I	37	36	24	3	100	40	33	25	2	100	39	35	27	3	104
	Total	80	62	53	9	204	86	65	47	6	204	83	61	55	9	208

Key: A = White; B = African American; C = Hispanic/Latino; O = Other Race.

Table 63

Frequencies by Group for Foundation Small Condition 2 (Method 1), Foundation Small Condition 2 (Method 2), and Foundation Small Condition 2-Randomized for Education

		FSC2				FSC2-2				FSC2-R			
Education	Group	A	B	C	Total	A	B	C	Total	A	B	C	Total
	C	38	31	30	99	33	36	30	99	45	31	24	100
	I	51	31	16	98	52	39	7	98	47	32	23	102
	Total	89	62	46	197	85	75	37	197	92	63	47	202

Key: A = Less than 12th grade; B = High School Diploma or GED; C = More than High School Education.

Table 64

Frequencies by Group for Foundation Small Condition 2 (Method 1), Foundation Small Condition 2 (Method 2), and Foundation Small Condition 2-Randomized for Primary Occupation

		FSC2				FSC2-2				FSC2-R			
Primary Occupation	Group	A	B	O	Total	A	B	O	Total	A	B	O	Total
	C	32	17	49	98	33	17	49	99	28	17	54	99
	I	20	23	55	98	20	26	52	98	18	29	54	101
	Total	52	40	104	196	53	43	101	197	46	46	108	200

Key: A = Employed; B = School or Training; O = Other primary occupation.

Table 65

Frequencies by Group for Foundation Small Condition 2 (Method 1), Foundation Small Condition 2 (Method 2), and Foundation Small Condition 2-Randomized for English Language Ability

		FSC2				FSC2-2				FSC2-R			
Language Ability	Group	A	B	C	Total	A	B	C	Total	A	B	C	Total
	C	80	3	16	99	86	4	9	99	79	5	15	99
	I	80	7	10	97	80	5	12	97	82	8	11	101
	Total	160	10	26	196	166	9	21	196	161	13	26	200

Key: A = Parent's primary language is English; B = Primary language is not English but the parent speaks English well; C = The primary language is not English and the parent does not speak English well.

Table 66

Frequencies by Group for Foundation Small Condition 2 (Method 1), Foundation Small Condition 2 (Method 2), and Foundation Small Condition 2-Randomized for Living Arrangement

		FSC2				FSC2-2				FSC2-R			
Living Arrange-ment	Group	A	B	C	Total	A	B	C	Total	A	B	C	Total
	C	34	37	35	106	28	39	39	106	29	37	40	106
	I	27	48	31	106	26	51	29	106	28	50	28	106
	Total	61	85	66	212	54	90	68	212	57	87	68	212

Key: A = Lives with husband; B = Lives with Other Adults; C = Lives alone.

Table 67

Frequencies by Group for Foundation Small Condition 2 (Method 1), Foundation Small Condition 2 (Method 2), and Foundation Small Condition 2-Randomized for Maternal Risk Index

		FSC2				FSC2-2				FSC2-R			
Maternal Risk	Group	A	B	C	Total	A	B	C	Total	A	B	C	Total
	C	51	36	4	91	47	27	19	93	40	32	20	92
	I	28	31	31	90	34	25	31	90	39	26	29	94
	Total	79	67	35	181	81	52	50	183	79	58	49	186

Key: A = 0, 1, or 2 risks; B = 3 risks; C = 4, 5 risks.

Table 68

Frequencies by Group for Foundation Small Condition 2 (Method 1), Foundation Small Condition 2 (Method 2), and Foundation Small Condition 2-Randomized for Child Age

		FSC2				FSC2-2				FSC2-R			
Child Age	Group	A	B	C	Total	A	B	C	Total	A	B	C	Total
	C	25	35	46	106	25	39	42	106	25	41	40	106
	I	27	42	37	106	25	41	40	106	25	41	40	106
	Total	52	77	83	212	50	80	82	212	50	82	80	212

Key (Child’s age at EHS application): A = Mother Pregnant; B = Child Less than 5 months old; C = Child More than 5 months old.

Table 69

Foundation Small Condition 2 (Method 1), Foundation Small Condition 2 (Method 2), and Foundation Small Condition 2-Randomized Balance Check: Independent Samples t-Tests (for Continuous Characteristics)

	FSC2			FSC2-2			FSC2-R		
	t	df	Sig. (2-tailed)	*t*	df	Sig. (2-tailed)	*t*	df	Sig. (2-tailed)
Mother's Age	1.335	209	.183	.612	209	.541	.183	209	.855
Child Gender	-1.514	210	.132	-.410	210	.682	-1.377	210	.170
Child Premature	-.603	145	.548	-.318	146	.751	-1.037	146	.301
Welfare	-.086	164	.931	1.222	163.524	.224	1.426	166	.156
Food Stamps	-.184	198	.854	1.205	198.648	.230	1.174	201	.242

Table 70

Foundation Small Condition 2 (Method 1), Foundation Small Condition 2 (Method 2), and Foundation Small Condition 2-Randomized Balance Check: Chi-squares Test of Independence (for Categorical Characteristics)

	FSC2			FSC2-2			FSC2-R		
	X^2	df	Sig. (2-sided)	X^2	df	Sig. (2-sided)	X^2	df	Sig. (2-sided)
Race	3.457	3	.326	1.214	3	.750	2.647	3	.449
Education	6.155	2	.046	18.660	2	.000	.061	2	.970
Primary Occupation	4.015	2	.134	5.157	2	.076	5.285	2	.071
Language Ability	2.965	2	.227	.736	2	.692	1.344	2	.511
Living Arrangement	2.469	2	.291	3.145	2	.208	4.078	2	.130
Maternal Risk Index	27.893	2	.000	4.996	2	.082	2.265	2	.322
Child's Age	1.689	2	.430	.099	2	.952	.000	2	1.000

Table 71

Foundation Small Condition 2 (Method 1) and Foundation Small Condition 2 (Method 2) Bias Check for Continuous (Comparing Disrupted to Randomized Standard, One Samples t*-tests)*

	FSC2						FSC2-2					
	n	Mean	SD	*t*	df	sig.	n	Mean	SD	*t*	df	sig.
Mother's Age	211	22.801	6.005	.321	210	.748	211	22.526	5.896	-.350	210	.727
Child Gender	212	.524	.501	-.411	211	.682	212	.514	.501	-.684	211	.494
Child Premature	147	.143	.351	.499	146	.618	148	.122	.328	-.251	147	.802
Welfare	166	.247	.433	-.090	165	.929	166	.265	.443	.438	165	.662
Food Stamps	200	.370	.484	-.561	199	.575	201	.408	.493	.540	200	.590

Table 72

Foundation Small Condition 2 (Method 1) and Foundation Small Condition 2 (Method 2) Bias Check for Categorical Variables

	FSC2				FSC2-2			
	n	X^2	df	Sig. (2-sided)	n	X^2	df	Sig. (2-sided)
Race	179	3.016	3	.389	204	4.774	3	.189
Education	172	15.282	2	.001	197	25.005	2	.000
Primary Occupation	171	2.891	2	.236	197	4.712	2	.095
Language Ability	172	2.072	2	.355	196	6.416	2	.040
Living Arrangement	187	8.861	2	.012	212	5.814	2	.055
Maternal Risk Index	156	42.605	2	.000	183	9.881	2	.007
Child's Age	212	2.187	2	.335	212	.198	2	.906

Noncompliance Conditions 10-18

Table 73

Means and Standard Deviations by Group for Foundation Small Condition 10 (Method 1), Foundation Small Condition 10 (Method -2), and Foundation Small Condition 10-Randomized on Five Continuous Baseline Characteristics

		FSC10			FSC10-2			FSC10-R		
	Group	n	Mean	SD	n	Mean	SD	n	Mean	SD
Mother Age	C	62	23.355	6.135	62	23.484	5.977	62	22.806	5.188
	I	62	22.742	6.583	62	22.161	6.438	61	22.918	6.515
Child Gender	C	62	.452	.502	62	.452	.502	62	.468	.503
	I	62	.565	.500	62	.500	.504	62	.629	.487
Child Premature	C	47	.064	.247	46	.130	.341	40	.175	.385
	I	42	.190	.397	45	.111	.318	47	.170	.380
Welfare Receipt	C	50	.300	.463	51	.294	.460	51	.314	.469
	I	47	.255	.441	49	.204	.407	50	.260	.443
Food Stamps Receipt	C	57	.439	.501	61	.410	.496	60	.433	.500
	I	60	.350	.481	58	.379	.489	60	.417	.497

Table 74

Frequencies by Group for Foundation Large Condition 10 (Method 1), Foundation Large Condition 10 (Method 2), and Foundation Large Condition 10-Randomized for Race

		FLC10					FLC10-2					FLC10-R				
Race	Group	A	B	C	O	Total	A	B	C	O	Total	A	B	C	O	Total
	C	23	16	18	4	61	24	19	16	2	61	27	14	18	3	62
	I	19	25	15	1	60	20	23	15	1	59	22	17	17	1	57
	Total	42	41	33	5	121	44	42	31	3	120	49	31	35	4	119

Key: A = White; B = African American; C = Hispanic/Latino; O = Other Race.

Table 75

Frequencies by Group for Foundation Large Condition 10 (Method 1), Foundation Large Condition 10 (Method 2), and Foundation Large Condition 10-Randomized for Education

		FLC10				FLC10-2				FLC10-R			
Education	Group	A	B	C	Total	A	B	C	Total	A	B	C	Total
	C	25	18	15	58	22	17	21	60	26	22	12	60
	I	30	16	13	59	27	17	14	58	26	21	9	56
	Total	55	34	28	117	49	34	35	118	52	43	21	116

Key: A = Less than 12th grade; B = High School Diploma or GED; C = More than High School Education.

Table 76

Frequencies by Group for Foundation Large Condition 10 (Method 1), Foundation Large Condition 10 (Method 2), and Foundation Large Condition 10-Randomized for Primary Occupation

		FLC10				FLC10-2				FLC10-R			
Primary Occupation	Group	A	B	O	Total	A	B	O	Total	A	B	O	Total
	C	17	8	33	58	24	9	26	59	19	7	33	59
	I	14	21	25	60	13	19	25	57	13	11	33	57
	Total	31	29	58	118	37	28	51	116	32	18	66	116

Key: A = Employed; B = School or Training; O = Other primary occupation.

Table 77

Frequencies by Group for Foundation Large Condition 10 (Method 1), Foundation Large Condition 10 (Method 2), and Foundation Large Condition 10-Randomized for English Language Ability

		FLC10				FLC10-2				FLC10-R			
Language Ability	Group	A	B	C	Total	A	B	C	Total	A	B	C	Total
	C	46	2	10	58	44	4	10	58	46	3	11	60
	I	49	4	6	59	50	6	2	58	42	7	7	56
	Total	95	6	16	117	94	10	12	116	88	10	18	116

Key: A = Parent's primary language is English; B = Primary language is not English but the parent speaks English well; C = The primary language is not English and the parent does not speak English well.

Table 78

Frequencies by Group for Foundation Large Condition 10 (Method 1), Foundation Large Condition 10 (Method 2), and Foundation Large Condition 10-Randomized for Living Arrangement

		FLC10				FLC10-2				FLC10-R			
Living Arrangement	Group	A	B	C	Total	A	B	C	Total	A	B	C	Total
	C	17	19	26	62	16	20	26	62	16	21	25	62
	I	13	31	18	62	13	30	19	62	17	29	16	62
	Total	30	50	44	124	29	50	45	124	33	50	41	124

Key: A = Lives with husband; B = Lives with Other Adults; C = Lives alone.

Table 79

Frequencies by Group for Foundation Large Condition 10 (Method 1), Foundation Large Condition 10 (Method 2), and Foundation Large Condition 10-Randomized for Maternal Risk Index

		FLC10				FLC10-2				FLC10-R			
Maternal Risk	Group	A	B	C	Total	A	B	C	Total	A	B	C	Total
	C	24	17	11	52	30	17	10	57	26	20	12	58
	I	22	15	17	54	23	16	15	54	20	13	18	51
	Total	46	32	28	106	53	33	25	111	46	33	30	109

Key: A = 0, 1, or 2 risks; B= 3 risks; C = 4, 5 risks.

Table 80

Frequencies by Group for Foundation Large Condition 10 (Method 1), Foundation Large Condition 10 (Method 2), and Foundation Large Condition 10-Randomized for Child Age

		FLC10				FLC10-2				FLC10-R			
Child Age	Group	A	B	C	Total	A	B	C	Total	A	B	C	Total
	C	12	24	26	62	13	24	25	62	17	21	24	62
	I	18	22	22	62	13	23	26	62	12	30	20	62
	Total	30	46	48	124	26	47	51	124	29	51	44	124

Key (Child's age at EHS application): A = Mother Pregnant; B = Child Less than 5 months old; C = Child More than 5 months old.

Table 81

Foundation Small Condition 10 (Method 1), Foundation Small Condition 10 (Method 2), and Foundation Small Condition 10-Randomized Balance Check: Independent Samples t-Tests (for Continuous Characteristics)

	FSC10			FSC10-2			FSC10-R		
	t	df	Sig. (2-tailed)	*t*	df	Sig. (2-tailed)	*t*	df	Sig. (2-tailed)
Mother's Age	.536	122	.593	1.185	122	.238	-.105	121	.916
Child Gender	-1.255	122	.212	-.536	122	.593	-1.814	122	.072
Child Premature	-1.780	67.082	.080	.280	89	.780	.058	85	.954
Welfare	.486	95	.628	1.037	97.353	.302	.592	99	.555
Food Stamps	.976	115	.331	.338	117	.736	.183	118	.855

Table 82

Foundation Small Condition 10 (Method 1), Foundation Small Condition 10 (Method 2), and Foundation Small Condition 10-Randomized Balance Check: Chi-squares Test of Independence (for Categorical Characteristics)

	FSC10			FSC10-2			FSC10-R		
	X^2	df	Sig. (2-sided)	X^2	df	Sig. (2-sided)	X^2	df	Sig. (2-sided)
Race	4.421	3	.219	1.077	3	.783	1.622	3	.654
Education	.707	2	.702	1.877	2	.391	.314	2	.855
Primary Occupation	7.190	2	.027	6.829	2	.033	1.980	2	.372
Language Ability	1.753	2	.416	6.116	2	.047	2.536	2	.281
Living Arrangement	4.868	2	.088	3.399	2	.183	3.286	2	.193
Maternal Risk Index	1.460	2	.482	1.875	2	.392	3.030	2	.220
Child's Age	1.620	2	.445	.041	2	.980	2.814	2	.245

Table 83

Foundation Small Condition 10 (Method 1) and Foundation Small Condition 10 (Method 2) Bias Check for Continuous (Comparing Disrupted to Randomized Standard, One Samples t*-Tests)*

	FSC10						FSC10-2					
	n	Mean	SD	*t*	df	sig.	*n*	Mean	SD	*t*	df	sig.
Mother's Age	124	23.048	6.345	.327	123	.744	124	22.823	6.222	-.070	123	.944
Child Gender	124	.508	.502	-.895	123	.373	124	.476	.501	-1.612	123	.110
Child Premature	89	.124	.331	-1.391	88	.168	91	.121	.328	-1.499	90	.137
Welfare	97	.278	.451	-.191	96	.849	100	.250	.435	-.852	99	.396
Food Stamps	117	.393	.491	-.702	116	.484	119	.395	.491	-.668	118	.506

Table 84

Foundation Small Condition 10 (Method 1) and Foundation Small Condition 10 (Method 2) Bias Check for Categorical Variables

	FSC10				FSC10-2			
	n	X^2	df	Sig. (2-sided)	n	X^2	df	Sig. (2-sided)
Race	121	5.621	3	.132	120	5.209	3	.157
Education	117	5.099	2	.078	118	12.080	2	.002
Primary Occupation	118	11.461	2	.003	116	11.130	2	.004
Language Ability	117	3.019	2	.221	116	5.749	2	.056
Living Arrangement	124	1.622	2	.444	124	1.626	2	.444
Maternal Risk Index	106	1.250	2	.535	111	3.041	2	.219
Child's Age	124	7.399	2	.025	124	4.928	2	.085

Differential Attrition Conditions

Table 85

Means and Standard Deviations by Group for Foundation Small Condition 19 (Method 1), Foundation Small Condition 19 (Method 2), and Foundation Small Condition 19-Randomized on Five Continuous Baseline Characteristics

		FSC19			FSC19-2			FSC19-R		
	Group	n	Mean	SD	n	Mean	SD	n	Mean	SD
Mother Age	C	62	22.468	5.288	62	22.919	6.165	62	22.806	5.188
	I	62	23.661	6.719	61	23.131	6.187	61	22.918	6.515
Child Gender	C	62	.387	.491	62	.468	.503	62	.468	.503
	I	62	.597	.495	62	.597	.495	62	.629	.487
Child Premature	C	41	.146	.358	45	.111	.318	40	.175	.385
	I	46	.196	.401	42	.190	.397	47	.170	.380
Welfare Receipt	C	52	.269	.448	49	.224	.422	51	.314	.469
	I	51	.216	.415	47	.213	.414	50	.260	.443
Food Stamps Receipt	C	61	.410	.496	58	.414	.497	60	.433	.500
	I	61	.393	.493	58	.431	.500	60	.417	.497
	I	62	22.468	5.288	62	22.919	6.165	62	22.806	5.188

Table 86

Frequencies by Group for Foundation Small Condition 19 (Method 1), Foundation Small Condition 19 (Method 2), and Foundation Small Condition 19-Randomized for Race

		FSC19					FSC19-2					FSC19-R				
Race	Group	A	B	C	O	Total	A	B	C	O	Total	A	B	C	O	Total
	C	25	17	16	2	60	25	15	18	2	60	27	14	18	3	62
	I	20	16	20	2	58	19	24	17	0	60	22	17	17	1	57
	Total	45	33	36	4	118	44	39	35	2	120	49	31	35	4	119

Key: A = White; B = African American; C = Hispanic/Latino; O = Other Race.

Table 87

Frequencies by Group for Foundation Small Condition 19 (Method 1), Foundation Small Condition 19 (Method 2), and Foundation Small Condition 19-Randomized for Education

		FSC19				FSC19-2				FSC19-R			
Education	Group	A	B	C	Total	A	B	C	Total	A	B	C	Total
	C	28	15	16	59	28	18	13	59	26	22	12	60
	I	27	20	10	57	30	17	13	60	26	21	9	56
	Total	55	35	26	116	58	35	26	119	52	43	21	116

Key: A = Less than 12th grade; B = High School Diploma or GED; C = More than High School Education.

Table 88

Frequencies by Group for Foundation Small Condition 19 (Method 1), Foundation Small Condition 19 (Method 2), and Foundation Small Condition 19-Randomized for Primary Occupation

		FSC19				FSC19-2				FSC19-R			
Primary Occupation	Group	A	B	O	Total	A	B	O	Total	A	B	O	Total
	C	19	13	27	59	19	9	31	59	19	7	33	59
	I	13	13	32	58	9	16	34	59	13	11	33	57
	Total	32	26	59	117	28	25	65	118	32	18	66	116

Key: A = Employed; B = School or Training; O = Other primary occupation

Table 89

Frequencies by Group for Foundation Small Condition 19 (Method 1), Foundation Small Condition 19 (Method 2), and Foundation Small Condition 19-Randomized for English Language Ability

		FSC19				FSC19-2				FSC19-R			
Language Ability	Group	A	B	C	Total	A	B	C	Total	A	B	C	Total
	C	45	5	7	57	42	4	11	57	46	3	11	60
	I	42	3	11	56	45	5	9	59	42	7	7	56
	Total	87	8	18	113	87	9	20	116	88	10	18	116

Key: A = Parent's primary language is English; B = Primary language is not English but the parent speaks English well; C = The primary language is not English and the parent does not speak English well.

Table 90

Frequencies by Group for Foundation Small Condition 19 (Method 1), Foundation Small Condition 19 (Method 2), and Foundation Small Condition 19-Randomized for Living Arrangement

		FSC19				FSC19-2				FSC19-R			
Living Arrangement	Group	A	B	C	Total	A	B	C	Total	A	B	C	Total
	C	14	24	24	62	16	22	24	62	16	21	25	62
	I	19	31	12	62	18	26	18	62	17	29	16	62
	Total	33	55	36	124	34	48	42	124	33	50	41	124

Key: A = Lives with husband; B = Lives with Other Adults; C = Lives alone.

Table 91

Frequencies by Group for Foundation Small Condition 19 (Method 1), Foundation Small Condition 19 (Method 2), and Foundation Small Condition 19-Randomized for Maternal Risk Index

		FSC19				FSC19-2				FSC19-R			
Maternal Risk	Group	A	B	C	Total	A	B	C	Total	A	B	C	Total
	C	26	15	16	57	23	21	10	54	26	20	12	58
	I	23	17	14	54	22	14	20	56	20	13	18	51
	Total	49	32	30	111	45	35	30	110	46	33	30	109

Key: A = 0, 1, or 2 risks; B= 3 risks; C = 4, 5 risks.

Table 92

Frequencies by Group for Foundation Small Condition 19 (Method 1), Foundation Small Condition 19 (Method 2), and Foundation Small Condition 19-Randomized for Child Age

		FSC19				FSC19-2				FSC19-R			
Child Age	Group	A	B	C	Total	A	B	C	Total	A	B	C	Total
	C	15	24	23	62	12	27	23	62	17	21	24	62
	I	14	26	22	62	17	19	26	62	12	30	20	62
	Total	29	50	45	124	29	46	49	124	29	51	44	124

Key (Child’s age at EHS application): A = Mother Pregnant; B = Child Less than 5 months old; C = Child More than 5 months old.

Table 93

Foundation Small Condition 19 (Method 1), Foundation Small Condition 19 (Method 2), and Foundation Small Condition 19-Randomized Balance Check: Independent Samples t-Tests (for Continuous Characteristics)

	FSC19			FSC19-2			FSC19-R		
	t	df	Sig. (2-tailed)	*t*	df	Sig. (2-tailed)	*t*	df	Sig. (2-tailed)
Mother's Age	-1.099	122	.274	-.190	121	.850	-.105	121	.916
Child Gender	-2.369	122	.019	-1.440	122	.152	-1.814	122	.072
Child Premature	-.602	85	.549	-1.024	78.492	.309	.058	85	.954
Welfare	.629	101	.531	.137	94	.891	.592	99	.555
Food Stamps	.183	120	.855	-.186	114	.852	.183	118	.855

Table 94

Foundation Small Condition 19 (Method 1), Foundation Small Condition 19 (Method 2), and Foundation Small Condition 19-Randomized Balance Check: Chi-squares Test of Independence (for Categorical Characteristics)

	FSC19			FSC19-2			FSC19-R		
	X^2	df	Sig. (2-sided)	X^2	df	Sig. (2-sided)	X^2	df	Sig. (2-sided)
Race	.997	3	.802	4.924	3	.177	1.622	3	.654
Education	2.083	2	.353	.089	2	.956	.314	2	.855
Primary Occupation	1.540	2	.463	5.670	2	.059	1.980	2	.372
Language Ability	1.484	2	.476	.380	2	.827	2.536	2	.281
Living Arrangement	5.648	2	.059	1.308	2	.520	3.286	2	.193
Maternal Risk Index	.361	2	.835	4.721	2	.094	3.030	2	.220
Child's Age	.137	2	.934	2.437	2	.296	2.814	2	.245

Table 95

Foundation Small Condition 19 (Method 1) and Foundation Small Condition 19 (Method 2) Bias Check for Continuous (Comparing Disrupted to Randomized Standard, One Samples t*-Tests)*

	FSC19						FSC19-2					
	n	Mean	SD	*t*	df	sig.	n	Mean	SD	*t*	df	sig.
Mother's Age	124	23.065	6.051	.373	123	.710	123	23.024	6.152	.293	122	.770
Child Gender	124	.492	.502	-1.253	123	.213	124	.532	.501	-.359	123	.720
Child Premature	87	.172	.380	.000	86	1.000	87	.149	.359	-.598	86	.552
Welfare	103	.243	.431	-1.045	102	.298	96	.219	.416	-1.612	95	.110
Food Stamps	122	.402	.492	-.524	121	.601	116	.422	.496	-.056	115	.955

Table 96

Foundation Small Condition 19 (Method 1) and Foundation Small Condition 19 (Method 2) Bias Check for Categorical Variables

	FSC19				FSC19-2			
	n	X^2	df	Sig. (2-sided)	n	X^2	df	Sig. (2-sided)
Race	118	3.117	3	.374	120	4.844	3	.184
Education	116	3.912	2	.141	119	4.120	2	.127
Primary Occupation	117	6.628	2	.036	118	4.226	2	.121
Language Ability	113	7.381	2	.025	116	2.038	2	.361
Living Arrangement	124	2.092	2	.351	124	.707	2	.702
Maternal Risk Index	111	5.153	2	.076	110	1.229	2	.541
Child's Age	118	3.117	2	.374	124	11.143	2	.004

APPENDIX D

SAMPLE PHASE III OUTCOME AND SENSITIVITY ANALYSIS

Condition 1 Phase III Data Tables

Table 97

Foundation Small Condition 1: Child Outcome Sample Sizes, Means, and Standard Deviations

		Threatened Method 1			Threatened Method 2			Randomized		
		N	Mean	SD	N	Mean	SD	N	Mean	SD
24m Child Engagement	control	73	4.51	1.00	75	4.36	1.18	79	4.33	1.15
	program	84	4.13	1.27	83	4.14	1.22	81	4.25	1.20
36m Child Engagement	control	71	4.65	1.04	72	4.79	1.10	73	4.68	1.05
	program	78	4.91	0.79	79	4.76	0.87	79	4.91	0.88
24m BSID-II MDI	control	75	90.13	12.84	74	90.62	13.24	78	90.97	12.21
	program	79	89.75	12.10	77	89.77	13.17	76	90.13	12.45
36m BSID-II MDI	control	72	90.10	13.94	72	90.69	14.10	75	91.17	11.64
	program	76	91.46	12.18	73	90.33	11.94	75	91.77	12.86
24m CBCL	control	86	22.35	11.23	87	22.65	11.53	87	22.02	10.49
	program	91	20.31	10.13	90	20.16	10.25	93	20.54	9.96
36m CBCL	control	89	19.57	11.32	89	20.49	11.92	91	18.73	10.65
	program	92	17.62	10.05	90	17.67	9.86	89	18.54	9.51
36m PPVT	control	62	81.79	16.70	68	82.10	17.15	65	81.98	16.19
	program	62	83.40	13.44	60	81.97	13.20	65	82.83	13.00

Table 98

Foundation Small Condition 1: Point-in-Time Outcomes (Intervention vs. Control Differences)

	Threatened Dataset 1				Threatened Dataset 2				Randomized Dataset			
	Test Value	df	Sig Level (α=.05)	Effect Size (Cohen's d)	Test Value	df	Sig Level (α=.05)	Effect Size (Cohen's d)	Test Value	df	Sig Level (α=.05)	Effect Size (Cohen's d)
24m Child Engagement	2.04	155	.04	0.33	1.12	156	.26	.18	.44	158	.66	-0.07
36m Child Engagement	-1.72	130	.09	-0.29	.20	149	.84	.03	-1.43	140	.15	0.23
24m BSID-II MDI	.19	152	.85	-0.03	.40	149	.69	.06	.42	152	.67	-0.07
36m BSID-II MDI	-.63	146	.53	0.10	.17	143	.87	.03	-.30	148	.76	0.05
24m CBCL	1.27	175	.20	-0.19	1.52	175	.13	.23	.97	178	.33	-0.14
36m CBCL	1.22	179	.22	-0.18	1.73	177	.09	.26	.13	178	.90	-0.02
36m PPVT	-.59	122	.55	0.11	.05	126	.96	.01	-.33	128	.74	0.06

Table 99

Foundation Small Condition 1: Sensitivity Analysis on Point-in-Time Results (Threatened vs. Randomized Differences)

Method 1						Method 2					
Error Type	Overlapping Effect Size C.I.s	ES Difference	Bias (SMD)	Percent Bias	Overall Biased	Error Type	Overlapping Effect Size C.I.s	ES Difference	Bias (SMD)	Percent Bias	Overall Biased
Type I or Type II	Y=yes, N=no	Criteria: |d| > .2	Criteria: |d| > .2	%>.05	Y=yes, N=no	Type I or Type II	Y=yes, N=no	Criteria: |d| > .2	Criteria: |d| > .2	%>.10	Y=yes, N=no
Type I	Y	-0.40	-0.25	-0.43%	Y	None	N	-0.25	0.03	0.95%	Y
None	Y	0.52	0.20	0.36%	Y	None	N	0.20	0.03	0.58%	Y
None	Y	-0.04	-0.13	0.69%	N	None	Y	-0.13	0.03	0.41%	N
None	Y	-0.06	0.02	0.74%	N	None	Y	0.02	0.08	1.05%	N
None	Y	0.05	-0.37	-0.22%	N	None	N	-0.37	-0.01	-0.59%	Y
None	N	0.16	-0.28	0.31%	N	None	N	-0.28	-0.04	-2.33%	Y
None	Y	-0.05	0.05	-0.23%	N	None	Y	0.05	0.02	0.45%	N

Table 100

Foundation Small Condition 1: Repeated Measures Analytic Results (Intervention vs. Control Differences)

	Method 1			Method 2			Randomized		
	Test Value	Sig Level	Effect Size (Cohen's d)	Test Value	Sig Level	Effect Size (Cohen's d)	Test Value	Sig Level	Effect Size (Cohen's d)
Child Engagement (within)	15.66	.00	.68	17.34	.00	.71	17.16	.00	.71
Child Engagement (between)	.01	.92	.02	.27	.61	.09	.81	.37	.15
BSID-II (within)	4.22	.04	.36	2.54	.11	.28	3.25	.07	.32
BSID (between)	.41	.52	.11	.01	.91	.02	.04	.85	.03
CBCL (within)	11.56	.00	.52	8.41	.00	.45	9.83	.00	.48
CBCL (between)	2.40	.12	.75	3.92	.05	.31	.45	.50	.10

Table 101

Foundation Small Condition 1: Sensitivity Analysis on Repeated Measures Results (Threatened vs. Randomized Differences)

		Method 1					Method 2		
Error Type	Overlapping Effect Size C.I.s	ES Difference	Bias	Overall Biased	Error Type	Overlapping Effect Size C.I.s	ES Difference	Bias	Overall Biased
Type I or Type II	Y=yes, N=no	Criteria: |d| > .2	Criteria: % > .10	Y=yes, N=no	Type I or Type II	Y=yes, N=no	Criteria: |d| > .2	Criteria: % > .10	Y=yes, N=no
None	Y	.02	0.85%	N	None	Y	.00	-3.73%	N
None	N	.14	110.93%	Y	None	Y	.07	157.28%	Y
Type I	Y	-.05	-17.48%	Y	None	Y	.03	4.98%	N
None	Y	-.08	-241.90%	N	None	Y	.01	39.82%	Y
None	Y	-.04	-11.60%	Y	None	Y	.03	7.87%	N
None	N	-.65	-146.87%	Y	Type I	Y	-.20	-228.44%	Y

APPENDIX E

PHASE IV EFFECTS OF STATISTICAL ADJUSTMENT PROCEDURES DATA TABLES

Table 102

Foundation Small Condition1 (Method 1): Child Outcome Sample Sizes, Means, and Standard Deviations

		PSA-Adjusted			Covariate-Adjusted		
		N	Mean	*SD*	*N*	Mean	*SD*
24m Child Engagement	control	133	4.48	0.96	73	4.51	1.00
	program	140	4.18	1.26	83	4.11	1.26
36m Child Engagement	control	127	4.80	1.03	71	4.65	1.04
	program	131	5.03	0.78	77	4.90	.79
24m BSID-II MDI	control	135	90.47	12.12	75	90.13	12.84
	program	125	90.47	12.68	78	89.51	11.99
36m BSID-II MDI	control	122	90.99	13.46	72	90.10	13.94
	program	126	91.80	11.85	75	91.07	11.77
24m CBCL	control	146	22.75	11.55	86	22.35	11.23
	program	153	20.47	10.60	90	20.26	10.17
36m CBCL	control	158	20.18	10.81	89	19.57	11.32
	program	153	17.89	11.01	91	17.60	10.11
36m PPVT	control	115	84.56	18.03	62	81.79	16.70
	program	97	85.16	13.85	61	83.05	13.26

Table 103

Foundation Small Condition 1 (Method 1): Point-in-Time Adjusted Outcome Estimates

	PSA-Adjusted Estimates				Covariate-Adjusted Estimates			
	Test Value (t)	df	Sig Level (α=.05)	Effect Size (Cohen's d)	Test Value (F)	df	Sig Level (α=.05)	Effect Size (Cohen's d)
24m Child Engagement	2.29	260	.023	-0.28	3.91	2	0.05	0.32
36m Child Engagement	-2.04	236	.042	0.26	2.66	2	0.10	0.27
24m BSID-II MDI	-.01	257	.995	0.00	0.07	2	0.79	0.04
36m BSID-II MDI	-.50	245	.615	0.06	0.17	2	0.68	0.07
24m CBCL	1.78	296	.076	-0.21	2.23	2	0.14	0.23
36m CBCL	1.85	309	.065	-0.21	1.82	2	0.18	0.20
36m PPVT	-.27	208	.788	0.04	0.29	2	0.59	0.10

Table 104

Foundation Small Condition 1 (Method 1): Repeated Measures Adjusted Outcome Estimates

	PSA-Adjusted Estimates			Covariate-Adjusted Estimates		
	Test Value (F)	Sig Level	Effect Size (Cohen's d)	Test Value (F)	Sig Level	Effect Size (Cohen's d)
Child Engagement (within)	60.94	.00	1.03	7.11	.01	.46
Child Engagement (between)	.23	.63	.06	.02	.88	.03
BSID-II (within)	14.69	.00	.54	3.20	.08	.32
BSID (between)	.86	.36	.13	.21	.65	.08
CBCL (within)	17.76	.00	.50	2.69	.10	.25
CBCL (between)	3.85	.05	.23	2.95	.09	.27

Table 105

Foundation Small Condition 1 (Method 1): Overall Propensity Score Analysis-Adjusted Results

	Bias Introduced in Threat Condition	Error Type	Assessment of Error Type	ES Difference	Assessment of ES Difference	Percent Bias Reduction	Percent Bias Assessment
	Y=yes, N=no	Type I or Type II	W=worse, B=better, NC=No change	[2]ESr-ESa	W=worse, B=better, NC=No change	pos=improved	Chg>10%
24m Child Engagement	Y	Type I	NC	-0.21	B	-0.46%	NC
36m Child Engagement	Y	Type I	W	-0.02	B	-2.06%	NC
24m BSID-II MDI	N	None	NC	0.07	NC	0.59%	NC
36m BSID-II MDI	N	None	NC	-0.02	NC	0.66%	NC
24m CBCL	N	None	NC	-0.06	NC	-1.32%	NC
36m CBCL	N	None	NC	-0.19	NC	-1.91%	NC
36m PPVT	N	None	NC	0.02	NC	-2.72%	NC
Child Engagement (within)	N	None	NC	-0.32	W	-20.00%	W
Child Engagement (between)	Y	None	NC	0.09	NC	-28.58%	W
BSID-II (within)	Y	Type I	NC	-0.22	W	-35.11%	W
BSID (between)	N	None	NC	-0.10	NC	-55.82%	W
CBCL (within)	Y	None	NC	-0.03	NC	1.67%	B
CBCL (between)	Y	Type I	W	-0.13	B	126.24%	B

Table 106

Foundation Small Condition 1 (Method 1): Overall Covariate-Adjusted Results

	Bias Introduced in Threat Condition	Error Type	Assessment of Error Type	ES Difference	Assessment of ES Difference	Percent Bias Reduction	Percent Bias Assessment
	Y=yes, N=no	Type I or Type II	W=worse, B=better, NC=No change	^{2}ESr-ESa	W=worse, B=better, NC=No change	pos=improved	Chg>10%
24m Child Engagement	Y	Type I	NC	-0.28	B	0.00%	NC
36m Child Engagement	Y	None	NC	-0.04	B	0.00%	NC
24m BSID-II MDI	N	None	NC	0.02	NC	0.00%	NC
36m BSID-II MDI	N	None	NC	-0.03	NC	0.00%	NC
24m CBCL	N	None	NC	-0.05	NC	0.00%	NC
36m CBCL	N	None	NC	-0.16	NC	0.00%	NC
36m PPVT	N	None	NC	-0.03	NC	0.00%	NC
Child Engagement (within)	N	None	NC	-0.32	W	-2.73%	NC
Child Engagement (between)	Y	None	NC	0.09	NC	-16.12%	W
BSID-II (within)	Y	None	B	-0.22	W	16.71%	B
BSID (between)	N	None	NC	-0.10	NC	100.07%	B
CBCL (within)	Y	Type II	W	-0.03	NC	10.65%	B
CBCL (between)	Y	None	B	-0.13	B	129.79%	B