A Series of Sensitivity Analyses Examining the What Works Clearinghouse’s Guidelines on Attrition Bias

A dissertation presented to

the faculty of

The Patton College of Education of Ohio University

In partial fulfillment

of the requirements for the degree

Doctor of Philosophy

Marsha S. Lewis

December 2013

© 2013 Marsha S. Lewis. All Rights Reserved.
This dissertation titled
A Series of Sensitivity Analyses Examining the What Works Clearinghouse’s Guidelines
on Attrition Bias

by
MARSHA S. LEWIS

has been approved for
the Department of Educational Studies
and the Patton College of Education by

Gordon P. Brooks
Professor of Educational Studies

Renée A. Middleton
Dean, The Patton College of Education
Abstract

LEWIS, MARSHA S., Ph.D., December 2013, Educational Research and Evaluation

A Series of Sensitivity Analyses Examining the What Works Clearinghouse’s Guidelines on Attrition Bias

Director of Dissertation: Gordon P. Brooks

This dissertation addresses the following overall research question: How do the amount and type of attrition, under varying assumptions of how much a subject’s likelihood of dropping out of a study is related to his or her outcome, impact randomized controlled studies by contributing to systematic bias? The study first replicates a study conducted on behalf of the U.S. Department of Education’s What Works Clearinghouse. Then, by applying a more systematic change in the magnitudes of the coefficients representing how much a subject’s likelihood of dropping out of the study is correlated to his or her outcome and also varying the differential attrition rates, the study helps address the question: How sensitive is the measure of bias to changes in attrition rates and/or the relationship between outcome and a participant’s propensity to respond in randomized controlled trials? The study also adds to the complexity of the bias modeling by addressing the question: How does varying the random error to simulate variations in the reliability of instruments used across studies impact the attrition thresholds?

The methodology consisted of a series eight of Monte Carlo simulations (50,000 replications each) programmed in R. Each simulation varied one or more of the following components: the relationship (or correlation) between the outcome at follow-up for a study participant and his or her propensity to respond (or likelihood of not attriting from
the study); the magnitude of differential attrition between the treatment and control
groups; and the random error generated by the reliability of the outcome instruments.

The sensitivity analyses indicate that the What Works Clearinghouse attrition bias
model is sensitive to changes in the assumptions about the relationship between attrition
and outcome. The patterns in the findings indicate that the difference in the relationship
between the propensity to respond and outcome in the model are as important to the bias
estimates as the overall and differential attrition. Modifying the attrition bias formula to
allow for the relationship between the propensity to respond and outcome to be less
impactful in the model until that relationship reaches a certain threshold overall
magnitude may help provide some more specific guidance to reviewers who may be
reviewing studies where, for example, they assume zero or near zero relationship of the
propensity to respond to outcome in the control group.

The conclusion reached by varying the random error term in the model in order to
address the potential impact of reliability on the bias thresholds indicates that the WWC
attrition bias thresholds may be somewhat sensitive to varying reliabilities of instruments
across studies. This sensitivity may necessitate the development of more specific
guidance for reviewers of certain types of studies for inclusion in the U.S. Department of
Education’s What Works Clearinghouse.
Dedication

I dedicate this completed dissertation to my son, Brent Lewis. Since you were born 14 years ago you have reminded me daily of what is really important, and that short list does not include a prestigious academic title. However, since I always tell you to do your best at everything you attempt and to always finish what you start, I had no choice but to complete this project.
Acknowledgements

I offer my sincerest heartfelt thanks to my graduate advisor and dissertation chair, Dr. Gordon Brooks, for his amazing programming skills, his general guidance, and mostly for his patience and humor over the many years of my journey.

I am forever grateful to my dissertation committee members, Dr. George Johanson, Dr. John Hitchcock, and Dr. Beth Vanderveer. Dr. Johanson’s vision helped shape an applied research question that would interest me and lead to my dissertation. Without his inquisitive nature, kindness, and patience, I could never have deciphered the original methodology nor would I have stuck with the work to completion. Dr. Hitchcock came to Ohio University at just the right time. He is one of the few people in the world to have experience with the subject of my work and understand the practical consequences of my research questions. Dr. Vanderveer stepped in at the dissertation defense stage and spent time to offer a fresh perspective in the final stages.

To my boss Mark Weinberg, my longtime research colleagues Ani Ruhil, Lesli Johnson, and Sara Boyd, my mentor Pat Dewees, and my friend Terry Murphy, the six of you always believed and reminded me that I would finally finish this task.

To my late parents, Fred Shook and Bonnie Nichols Shook, you were the first two in your large families to have been given the luxury of graduating from high school, and you made sure I had the luxury of going to college. Thank you both for the incredibly rich gift of growing up among the many scholars in Garfield Hollow.

Lastly and always, I am thankful that God’s mercies are new every morning (Lamentations 3:22-23).
# Table of Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>3</td>
</tr>
<tr>
<td>Dedication</td>
<td>5</td>
</tr>
<tr>
<td>Acknowledgements</td>
<td>6</td>
</tr>
<tr>
<td>List of Tables</td>
<td>10</td>
</tr>
<tr>
<td>List of Figures</td>
<td>12</td>
</tr>
<tr>
<td>Chapter 1: Introduction</td>
<td>13</td>
</tr>
<tr>
<td>Background of the Study</td>
<td>13</td>
</tr>
<tr>
<td>Statement of the Problem</td>
<td>20</td>
</tr>
<tr>
<td>Subproblems</td>
<td>20</td>
</tr>
<tr>
<td>Significance of the Study</td>
<td>21</td>
</tr>
<tr>
<td>Policy Implications</td>
<td>22</td>
</tr>
<tr>
<td>Research Value</td>
<td>24</td>
</tr>
<tr>
<td>Delimitations and Limitations of the Study</td>
<td>25</td>
</tr>
<tr>
<td>Definition of Terms</td>
<td>27</td>
</tr>
<tr>
<td>Chapter 2: Review of the Literature</td>
<td>30</td>
</tr>
<tr>
<td>The Current Policy Context for Education Research</td>
<td>30</td>
</tr>
<tr>
<td>Definition and Purpose of Randomized Controlled Trials</td>
<td>32</td>
</tr>
<tr>
<td>The Call for Randomized Controlled Trials in Education Research</td>
<td>35</td>
</tr>
<tr>
<td>The What Works Clearinghouse Procedures and Standards Handbook</td>
<td>39</td>
</tr>
<tr>
<td>Critiques of the What Works Clearinghouse</td>
<td>40</td>
</tr>
</tbody>
</table>
Replicating and Extending What Works Clearinghouse Attrition Bias Model ............. 90

Chapter 5: Discussion, Conclusions, and Recommendations ............................................. 106

Discussion of the Results ........................................................................................................ 108

Conclusions ............................................................................................................................... 113

Limitations ................................................................................................................................. 116

Recommendations ..................................................................................................................... 117

References .................................................................................................................................. 122

Appendix A: Scenario 1 R Code Replicating Table A1 from What Works Clearinghouse
Procedures and Standards Handbook .......................................................................................... 131

Appendix B: Scenario 2 R Code Modeling the Counterfactual ................................................. 134

Appendix C: Scenario 3 R Code Systematically Extending Alphas and Response Rates
for Sensitivity Analysis ................................................................................................................. 137

Appendix D: Scenario 4 R Code Systematically Extending Alphas and Response Rates
for Sensitivity Analysis, Part 2 ................................................................................................... 140

Appendix E: Scenario 5 R Code Setting Control Group Relationship of Response to
Outcome at Zero ............................................................................................................................ 143

Appendix F: Scenario 6 R Code Holding Control Group Relationship of Response to
Outcome at Zero and Holding Control Group Attrition Rate at 10 Percent ......................... 146

Appendix G: Scenario 7 R Code Setting the Participant Group and Control Group
Relationship of Response to Outcome as Equal ......................................................................... 149

Appendix H: Scenario 8 R Code Simulating Sensitivity of Model to Variation in
Reliability of Outcome Instrument Across Studies .................................................................. 152
List of Tables

Table 1. What Works Clearinghouse’s Modeled Scenarios for Proportion of Outcome Explained by Response .................................................................18

Table 2. What Works Clearinghouse Simulation Results for Bias by Response Rate and Proportion of Outcome Explained by Response ...............61

Table 3. What Works Clearinghouse Simulation Results Illustrating Overall Differences between Respondents and Nonrespondents and the Difference in That Difference between the Treatment and Control Groups in the Case Of 90% Response and No Differential Attrition.............................................65

Table 4. Results of Scenario 1 Simulation: Replicating Original Attrition Bias Model ...........................................................................................92

Table 5. Results of Scenario 2 Simulation: Modeling the Counterfactual .................94

Table 6. Results of Scenario 3 Simulation: Systematically Extending Alphas for Sensitivity Analysis........................................................................96

Table 7. Results of Scenario 4 Simulation: Systematically Extending Alphas for Sensitivity Analysis, Part 2 .................................................................97

Table 8. Results of Scenario 5 Simulation: Setting Control Group Relationship of Response to Outcome at Zero .................................................................99

Table 9. Results of Scenario 6 Simulation: Holding Control Group Relationship of Response and Outcome to Zero and Holding Control Group Attrition Rate At 10 Percent........................................................................................................101
Table 10. Results of Simulation Scenario 7: Setting the Participant Group and Control Group relationship of Response to Outcome as Equal ..................103

Table 11. Results of Simulation Scenario 8: Simulating Sensitivity of Model to Variation in reliability of Outcome Instrument Across Studies......................105
### List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>WWC Illustration of Trade-Offs Between Overall and Differential Attrition</td>
<td>58</td>
</tr>
</tbody>
</table>
Chapter 1: Introduction

Background of the Study

In experimental research, attrition occurs when participants drop out of a study, making it impossible in many cases for researchers to collect outcome measures on all participants who were originally assigned to a treatment or control group (Shadish, Cook, & Campbell, 2002). Attrition occurs for a variety of reasons, sometimes dependent on the length of time that participants must stay in the study, the intrusiveness of the treatment or measurement, or characteristics of the study sample. Participants may drop out of longitudinal studies before all of the treatment has been delivered, or participants may simply refuse or be unavailable for outcome data collection even after the treatment has been administered (Enders, 2010; Puma, Olsen, Bell, & Price, 2009; Schafer & Graham, 2002).

For example, a study examining the success of college access programming in high schools that attempts to follow participants after high school graduation will likely suffer attrition because it is difficult to track students who may have moved or changed contact information during the course of the study. Also, students in the college access program who did not follow through with their plans to go on to college may not want to admit that they did not do what they said they were going to, so they choose not to respond to follow-up surveys that are designed to track outcomes. Study participants from families with lower incomes may be more transient than their higher income counterparts, and therefore more difficult to locate over the course of the study. These are just some of many plausible reasons why attrition may occur in this one study alone.
Unless a researcher is deploying an intention-to-treat analysis (Gupta, 2011) or has the ability to obtain outcome measures even after a participant drops out of a study, attrition almost always results in missing data. However, not all missing data are the same. There are conventions for classifying and describing missing data. These categories are defined largely by the relationship of the missing data to other observed and unobserved variables. Rubin (1976) first developed and labeled the categories that are still widely used to describe the context of the “missingness.” These three categories and their common acronyms are: Missing Completely At Random (MCAR), Missing At Random (MAR), and Missing Not At Random (MNAR). With data missing completely at random (MCAR), the loss of data is not dependent on or related to any observed or unobserved variable. For example, an employee’s 6-month evaluation rating score is missing from a dataset of employee job performance ratings because she moved before her 6-month evaluation was completed when her husband got a job in another state for reasons unrelated to her employment (Enders, 2010). Data missing at random (MAR) has the characteristic of the missing data being related to one or more of the observed variables, but not on the outcome variable itself. For example, a researcher may find that Hispanic students had more missing reading test scores than Caucasian students in a study of reading achievement (Enders, 2010). Data missing not at random (MNAR) has the characteristic of the missing data being dependent on the outcome variable itself. For example, students with poor reading skills have missing reading test scores. It is usually impossible to prove that data are MAR as opposed to MCAR; this can only be assumed (Enders, 2010; Schafer & Graham, 2002).
Another way to categorize and describe attrition relates to the difference in the amount of missing outcome data between the treatment and control groups, or differential attrition. Differential attrition that is MNAR is the most problematic for researchers because it poses a serious threat to the validity of the research findings (Puma et al., 2002; Shadish, Hu, Glaser, Kownacki, & Wong, 1998; Valentine & McHugh, 2007). For example, if students in a with lower reading skills are more likely to having missing outcome data for a reading intervention program than students in the control group, then the assumption that the treatment group and the control group are similar is no longer appropriate.

Any kind of attrition is unwanted, but attrition in some circumstances is more harmful to a study than in others. Even when attrition is random, statistical power is affected by the loss of cases for analysis. However, nonrandom attrition, or missing data caused by attrition that are MNAR, is a threat to the validity of a study’s findings (Valentine & McHugh, 2007). In order for researchers running randomized, controlled trials (RCTs) to get unbiased estimates of the impact of the treatment on the outcome, the treatment and control groups must be equivalent not only at the point of randomization, but also at follow-up. If not, then non-random differences may exist between the two groups, biasing the estimates of the effect of the treatment (Puma et al., 2009). The major advantage of a RCT is that the treatment and control groups are generally equivalent with any imbalance occurring by chance alone. Any loss to follow-up prevents a full analysis of the treatment’s impact on outcome and may introduce bias (Dumville, Torgerson, & Hewitt, 2006). Shadish, Cook, and Campbell (2002) listed attrition as one of the nine
threats to the internal validity of a study. They warned that the benefits of random assignment are compromised by attrition, which they indicated can rarely be assumed to be random, because the equivalence of the groups that was present at pretest cannot be assumed to carry over to posttest.

For most fields of inquiry, there has been little guidance for researchers on what levels of overall and differential attrition are acceptable in experimental research. Arbitrary rules of thumb have been set at 20 percent or less of overall attrition as acceptable (Amico, 2009; Schultz & Grimes, 2002; Valentine & McHugh, 2007). Methodologists who have studied the effects of attrition argue that these general thresholds oversimplify the attrition problem by ignoring such important factors as the context in which attrition occurs and the differing rates of attrition among the treatment levels (Amico, 2009; Valentine & McHugh, 2007). As Hewett, Kumaravel, Dumville, and Torgerson (2010) stated, “Potentially, 5% nonrandom attrition on an important prognostic variable could introduce more bias than 20% random attrition.” (p. 1265)

Probably the most sophisticated guidance to date on dealing with attrition in experimental designs comes from the U.S. Department of Education’s What Works Clearinghouse (WWC) Procedures and Standards Handbook, Version 2, published in 2008 and updated to Version 2.1 in 2011. (A new version, 3.0 has been developed and is currently under review but the official handbook is still Version 2.1. Version 3.0 has some minor changes to the attrition bias estimation procedure, but the basic structure remains the same.) The handbook describes the criteria and procedures that WWC reviewers use to screen studies and determine if they are worthy of admission into the
Clearinghouse. This review process is intentionally rigorous, as the WWC’s mission is to be a “central and trusted source of scientific evidence for what works in education.” (U.S. Department of Education, 2008, p. 1) The WWC does not conduct the research, but reviews and synthesizes findings from existing research in order to inform educators about the effectiveness of an intervention. Research that does not meet the strict evidence standards set by the WWC does not get included in the reviews of particular interventions. The amount and type of attrition in a study presented for WWC review is a key determinant of whether that study gets included in the Clearinghouse (U.S. Department of Education, 2008, 2011b).

In its Procedures and Standards Handbook, the WWC provides guidance on what is “acceptable,” “potentially acceptable,” and “unacceptable” levels of overall and differential attrition for experimental studies reviewed by the U.S. Department of Education and potentially added to the Clearinghouse (U.S. Department of Education, 2011, p. 13). The criteria were developed through a simulation study that, based on an arbitrary threshold of 0.05 standard deviation impact on the outcome measure as an acceptable level of bias, modeled the amount of both overall and differential attrition that could hold the bias to that impact or less. The simulation modeled the overall and differential attrition rates under different assumptions of the correlation between the propensity to respond (attrition), and the outcome at follow up.

Specifically, the researchers modeled eight conditions of correlation between whether or not a person responded and their outcome (proportion of outcome explained by the propensity to respond), one where the correlation between the propensity to
respond and outcome was the same for both the treatment and control groups and seven
where the relationship (expressed as a correlation) between the propensity to respond
and outcome differed between the treatment and control groups. Table 1 illustrates the
eight conditions of proportion of outcome explained by the propensity to respond used in
the WWC simulation.

Table 1
What Works Clearinghouse’s Modeled Scenarios for Proportion of Outcome Explained
by the Propensity to Respond (U.S. Department of Education, 2008)

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
<th>(7)</th>
<th>(8)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment Group</td>
<td>$\alpha_t=0.075$</td>
<td>$\alpha_t=0.10$</td>
<td>$\alpha_t=0.15$</td>
<td>$\alpha_t=0.20$</td>
<td>$\alpha_t=0.30$</td>
<td>$\alpha_t=0.50$</td>
<td>$\alpha_t=1.00$</td>
<td>$\alpha_t=1.00$</td>
</tr>
<tr>
<td>Control Group</td>
<td>$\alpha_c=0.05$</td>
<td>$\alpha_c=0.05$</td>
<td>$\alpha_c=0.05$</td>
<td>$\alpha_c=0.15$</td>
<td>$\alpha_c=0.20$</td>
<td>$\alpha_c=0.20$</td>
<td>$\alpha_c=1.00$</td>
<td>$\alpha_c=-1.00$</td>
</tr>
</tbody>
</table>

Of course, it is impossible to know the outcome for nonrespondents, so the
correlation models a theoretical relationship. Clearinghouse researchers attempted to
gauge the plausibility of the theoretical correlation by using an existing study and
students’ pretest scores as proxies for what their outcome would have been had they not

Building on the modeling done by the WWC researchers, Nianbo Dong and Mark
W. Lipsey of Vanderbilt’s Peabody Research Institute replicated the original simulation
study and expanded it to model the impact of adding a baseline covariate to control for
some of the bias, as well as applying the model to cluster randomized experiments (Dong
& Lipsey, 2010). Dong and Lipsey stressed that the WWC work on attrition bias was a
major contribution to the field: “Remarkably, given how much awareness there is in the experimental research community of the problems associated with attrition, little systematic research has been done to provide a framework for appraising attrition and assessing its potential to bias effect estimates.” (Dong & Lipsey, 2010, p. 2)

The WWC model provides that framework for reviewing educational research and making specific operational and policy decisions at the federal, state, and local levels. Decisions such as whether or not to fund and promote a particular reading program federally or purchase a costly educational software application locally may be driven largely by the rating given to a study by WWC reviewers. The bias thresholds in the WWC Procedures and Standards Handbook can be thought of as the “context-specific precision” articulated by Kane (1996). The bias estimates are being used to make consequential and often costly policy and practice decisions and draw conclusions, so they should be as precise as possible.

The WWC guidance for researchers on acceptable ranges of overall and differential attrition is important work, as it guides researchers and consumers of research, including policymakers, as they decide whether the findings of a particular study are valid. This simulation study adds value to the findings of the WWC simulation by more systematically modeling additional relationships between attrition, data missing MNAR, and outcomes and introducing the impact of test reliability that may vary from study to study as a component of the error term in the model. It, like Dong and Lipsey’s (2010) study, replicates the modeling and advances the research by introducing new components.
Statement of the Problem

This simulation provides additional data to address the following research question: How do the amount and type of attrition, under varying assumptions of how much subjects’ likelihood of dropping out is related to their outcome (MNAR), impact RCTs by contributing to systematic bias?

Subproblems. By applying a finer and more systematic change in the magnitudes of the coefficients representing how much a subject’s likelihood of dropping out of the study is correlated to his or her outcome and also varying the differential attrition rates, the study provides a more detailed answer to the question: How sensitive is the measure of bias to changes in attrition rates and/or the relationship between outcome and a participant’s propensity to respond in RCTs?

The study also adds to the complexity of the bias modeling originally developed by the WWC by addressing the question: How does varying the random error to simulate variations in the reliability of instruments used across studies impact the attrition thresholds? For example, if an outcome measure in one study has a reliability of .50 and the outcome measure in a second study has a reliability of .85, there is more random error inserted by the reliability of the instrument in the first study. This study models a scenario where more random error is inserted to simulate a lower reliability of an outcome instrument.

Finally, the study uses data visualization to help address the question: Are there discernible patterns to the increase or decrease in bias at different increments of overall and differential attrition and propensity to respond in RCTs?
Significance of the Study

Without evidence to the contrary, researchers must assume that attrition is systematic and therefore biases any results (Shadish, et al., 2002). In an article titled, “Percent Total Attrition: A Poor Metric for Study Rigor in Hosted Intervention Designs,” health researcher K. Rivet Amico (2009) stated,

Methodological rigor in behavioral intervention research is essential for the systematic identification of interventions that are empirically demonstrated to be effective and…appropriate for widespread dissemination and adoption. The establishment and implementation of formulas for defining rigor, though essential in this process, must nonetheless demonstrate sensitivity to a wide diversity of implementation and evaluation strategies. The manner in which attrition and retention factor into these formulas is an important, though understudied, consideration, which can exert influence on policy development. Oversimplified application of attrition as exclusion criteria can produce public policy that is biased toward certain types of research and away from others. (p. 1)

Valentine and McHugh (2007) conducted a meta-analysis examining the effects of attrition on baseline comparability in randomized experiments in education. Of the 367 articles selected for the analysis, 119 of the studies had clearly experienced attrition. Of the 119 studies that experienced attrition, only 26 provided sufficient information allowing for a calculation of an effect size of baseline comparability. They looked closely at these 26 studies and found the attrition rate averaged 16 percent, with some of the studies having rates as high as 30 percent. They conclude that “much greater attention
is needed to both the reporting of attrition in primary studies and to the development of conceptual and empirical models of the attrition process.” (Valentine & McHugh, 2007, p. 268)

This study follows Valentine and McHugh’s conclusion by further developing an empirical model of the attrition process. It provides additional detail about bias based on the amount and type of attrition experienced in any study. Like the Dong and Lipsey (2010) work, this study first replicates the findings of the WWC researchers, and then expands the simulation. It also provides visual models that aid in the interpretation of the results and the identification bias sensitivity patterns under various conditions.

By varying the random error term used in the original WWC simulation in order to account for the fact that outcome instrument reliability may be relatively large in some studies, the study provides information to researchers and policymakers regarding conditions that may contribute to bias.

**Policy implications.** As part of sweeping and unprecedented educational reforms instituted by the 2001 reauthorization of the Elementary and Secondary Education Act (Beghetto, 2003), the WWC was created in hopes to increase the rigor of educational research and provide educators a source for quality research on practices to improve student achievement. As indicated on the WWC website, one goal is to “connect educators with the best research on effective interventions and practices in education.” Educators are encouraged to use the site to help make decisions about programs, product purchases, and policies that improve student outcomes (Mageau, 2004). Slavin (2008)
stated, “Potentially, the WWC is the most important of the synthesis efforts for policy, because it alone carries the endorsement of the U.S. Department of Education.” (p. 6)

Because of the significant impact of the WWC on the educational research landscape since 2002, driving the field toward scientifically-based research/RCTs, the parameters set for reviews of studies seeking to be included in the clearinghouse are themselves important and have implications for policy, funding decisions, and the direction of educational interventions. School leaders are at least guided and at most required to select products and practices based on whether or not they are scientifically-based (Beghetto, 2003; Puma et al., 2009; Zucker, 2004). As of 2009, 25 RCTs examining educational interventions were being conducted in the federal government’s Regional Educational Laboratories and sponsored by the U.S. Department of Education (Puma et al., 2009). Indeed the call for RCTs is having the desired effect of increasing the number of these studies being conducted by educational researchers. These studies, in turn, have the potential to influence policy, purchasing, and even intervention strategies for individual students.

Because of this change in federal policy regarding what kind of educational research gets funded, what used to be confined to debates among academicians on topics such as effect sizes and attrition bias thresholds are now operationalized by reviewers who get to decide what evidence makes it into the clearinghouse and what does not. Therefore, the admittedly arbitrary thresholds of attrition bias (U.S. Department of Education, 2008) are impacting educational policy and practice. Since the U.S. Department of Education uses an attrition threshold as one of the key rating criteria for
stating that a study “Meets Evidence Standards” (U.S. Department of Education, 2011b, p. 11), work that substantiates and adds additional information to the calculations of bias caused by attrition can only strengthen the rigor and quality of educational research, which in turn is influencing what research gets funded, what educational interventions are marketable, and in turn, what teachers and students do in the classroom.

Since its development almost a decade ago, the WWC has been criticized by some in the educational research community for its overreliance on one type of research, its seemingly arbitrary decisions on some studies that the critics felt were judged either too leniently or with a cursory review that missed some key methodological information that could have changed the WWC rating (Confrey, 2006; Greene, 2010; Schoenfeld, A. H., 2006). Therefore, works such as this study that serve to cross-validate, illuminate, and add additional value to the work of the WWC designers is important to the field of educational research today and into the foreseeable future, as the WWC has survived a change in Presidential administrations and political parties and continues to be the official federal arbiter of what educational interventions should be seen as effective (U.S. Department of Education, 2011a).

**Research value.** In addition to the substantive value of the information on the sensitivity of bias to attrition, this study also, by providing all of the programming code with annotations, provides a way for researchers to easily replicate the findings and test other attrition conditions. Neither the WWC researchers nor Dong and Lipsey have provided their programming code to date. And, while Dong and Lipsey replicated the
findings using SAS; this study uses R (the same statistical package used by the WWC researchers) which is free and publicly available.

Howard Bloom, at the end of his 2006 MDRC working paper on the core analytics of randomized experiments for social research, stated that:

Perhaps the most important frontier for randomized experiments in the social sciences is the much-needed expansion of organizational and scientific capacity to implement them successfully on a much broader scale. To conduct this type of research well requires high levels of scientific and professional expertise, which at present exist only at a limited number of institutions. (p. 19)

The goal of this study on attrition bias is to add a small but important component to the scientific and professional expertise regarding RCTs and other research designs such as single case designs and regression discontinuity designs that have been added to the list of acceptable designs that meet the WWC standard (U.S. Department of Education, 2011b).

**Delimitations and Limitations of the Study**

The scope of this study is limited to: 1) replicating the WWC findings on the amount and type of attrition, under varying assumptions, that yields an acceptable amount of bias; 2) adding to the simulation additional conditions of the correlation between whether or not a person responds and his or her outcome score as well as varying the rates of overall and differential attrition; 3) varying the original simulation algorithm to model variations across studies in the reliability of the instrument that measures the outcome variable to the model as a portion of the error term in order to gauge the
potential impact of reliability on the acceptable thresholds of overall and differential attrition; and 4) using data visualization to illustrate the sensitivity of the bias statistic to changes in attrition and the relationship between the propensity to respond and outcome. The study does not replicate or advance the Dong and Lipsey (2010) study that adds a covariate to the model and examines the effect of attrition in cluster RCTs.

**Limitations.** As with any simulation study, it is impossible to know with absolute certainty that the assumptions and findings are plausible. For example, as stated in the WWC guide, we rarely have the opportunity to observe actual outcomes of nonrespondents, so there is no way to truly gauge whether the assumptions about the extent to which response and outcome are correlated are plausible assumptions (U.S. Department of Education, 2008).

As reported by Dong and Lipsey (2010):

…the results heavily depend on the model of attrition. The model of attrition used in this study assumes a linear relationship between the outcome and the propensity to respond. In addition, the propensity to respond is assumed a normal distribution, in which attrition occurs when the propensity to respond is under a certain threshold. These assumptions could be over-simplified and could not represent the real attrition mechanism. (pp. 49-50)

**Delimitations.** This study is only identifying and manipulating a particular set of variables that potentially introduce systematic bias in RCTs, including a correlation term representing the relationship between the propensity to respond and the individual’s outcome at follow up, the amount of overall attrition, the ratio of differential attrition, and
the error introduced by the reliability of the outcome instrument. Other variables that may introduce systematic bias were not considered or modeled.

**Definition of Terms**

1) *Attrition* - Attrition is the loss of study participants after random assignment has taken place. Measurement attrition is a type of attrition that indicated the failure to obtain outcome measures on participants, whether or not they were treated (Shadish, et al., 2002).

2) *Bias* - Bias is the systematic error in an estimate (Shadish et al., 2002).

3) *Control Group* - In experimental research, the control group is the group that does not receive the treatment (Shadish et al., 2002).

4) *Differential Attrition* - Differential attrition is the difference between the rate of attrition in the treatment group and the rate of attrition in the control group for a RCT (U.S. Department of Education, 2011b).

5) *External Validity* - Study findings about causal relationships possess external validity if the findings can be generalized to other persons and settings (Shadish et al., 2002).

6) *Internal Validity* - Studies possess internal validity if a causal relationship between two variables is properly demonstrated (Shadish et al., 2002).

7) *Missing At Random (MAR) data* - Missing data when the probability of data being missing on the outcome variable is related to other measured variables in the study but not to the outcome variable itself (Enders, 2010)
8) *Missing Completely at Random (MCAR) data*- “Purely haphazard missingness” where the probability of data being missing on the outcome variable is unrelated to any other measured variable in the study and to the outcome variable itself (Enders, 2010, p. 7).

9) *Missing Not at Random (MNAR) data*- Missing data when the probability of data being missing on the outcome variable is related to the outcome variable itself (Enders, 2010).

10) *Overall attrition*- The total attrition in the study across the treatment and control groups combined is the overall attrition.

11) *Propensity to Respond*- This is a term used in the What Works Clearinghouse Procedures and Standards Handbook Version 2 to refer to a subject’s likelihood of responding to outcome data collection, or their likelihood of not dropping out of the study before outcome data can be collected (U.S. Department of Education, 2008). The term has no relationship to propensity score matching.

12) *Random Assignment*- Random assignment is the procedure for assigning participants to treatment or control groups based on chance, with every participant having a nonzero probability of being assigned to each group (Shadish et al., 2002).

13) *Randomized Controlled Trial (RCT)*- A type of study in which researchers randomly assign eligible subjects into groups to receive or not receive one or more interventions that are being compared is known as a Randomized Controlled Trial.
14) *Reliability* - Reliability is the measure of the consistency or reproducibility of a test score (Crocker & Algina, 1986).
Chapter 2: Review of the Literature

The literature review is divided into five sections. The first section provides an overview of the current policy context for education research. The second section provides definition and purpose of RCTs. Section 3 describes the call for RCTs in education research. Section 4 specifically describes attrition as a threat to the validity of research findings in RCTs. The last section discusses measurement error as a threat to the validity of research findings in RCTs.

The Current Policy Context for Education Research

In his plenary address at the 2009 Institute of Education Sciences (IES) Research Conference, U.S. Education Secretary Arne Duncan articulated the Obama Administration’s challenge to the educational research community:

Reforming public education is not just a moral obligation. It is an absolutely and [sic] economic imperative. It is the foundation for a strong future and a strong society. Education is the civil rights issue of our generation. The fight for quality education is about so much more than education. It's a fight for social justice. It is the only way to achieve the quality that inspired our democracy that inspired women to stand up for their rights, and then inspired minorities to demand their fair share of the American promise and it inspires every child to dream…Those dreams are shaped in America's classrooms. They are nurtured by the dedicated teachers and principals all across America who do the hard work every single day of educating our children. And they are counting on all of you to help them get better, help them see how they can improve and help them turn their students'
dreams into reality…Education reform is not about sweeping mandates or grand gestures. It's about systematically examining and learning and building on what we're doing right and scrapping what hasn't worked for our children. (Duncan, 2009, p. 1)

This clarion call to broad public values, delivered to a room largely full of statisticians who were about to sit through a series of presentations on more granular topics as effect sizes and latent variables, is indicative of the change in the educational research community in recent years—the elevated importance of the work being done by academicians and private sector think tanks to figure out “what works” in the classroom. Words like “moral obligation,” “economic imperative,” and “social justice” and “systematic examination” were likely chosen carefully by Duncan, who is charged with leading one of the most dramatic reform periods in the history of U.S. public education.

The Secretary of Education was clear about the importance of educational research in the current reform efforts. The policymakers need these researchers in order to forward their goals, as the reform has been defined as research driven (U.S. Department of Education, 2011a). Topics such as effect sizes and standard errors used to be discussed almost exclusively among researchers and statisticians. These terms are now being discussed by teachers and principals and have a direct and immediate impact on policy and practice as well as on the evaluation of individual teachers, as the U.S. Department of Education is requiring states to measure teacher and principal performance by the academic growth of their students as measured by value-added models (U.S. Government Accountability Office, 2013).
Attrition bias is another term heretofore familiar only to a relative handful of quantitative researchers across the various disciplines, but now one of the key arbiters of whether or not the results of an educational study should be taken seriously by practitioners and policymakers and whether or not an educational intervention should be labeled as evidence-based and recommended for larger-scale deployment. Attrition bias is a special subset of selection bias that occurs when subjects drop out or do not complete the outcome measure after they have been assigned to either the treatment or control groups in an experiment (Shadish et al., 2002). Attrition bias is one of the handful of “reasons to think that the relationship between A and B is not causal” (Shadish et al., 2002, p. 54). Attrition bias is arguably the one condition that, if present and of a certain magnitude and type in an experimental study, automatically erodes the experimental design into a quasi-experimental design at best (Yeaton, Wortman, & Langberg, 1983). It is the magnitude and type of attrition and its impact on bias that is the focus of this study.

**Definition and Purpose of Randomized Controlled Trials**

A randomized experiment or randomized controlled trial (RCT) is a research design in which individuals or units are assigned to the treatments being studied, or to a control group not receiving the treatment, completely by chance. (Bloom, 2005, 2006; Puma et al., 2009; Shadish et al., 2002). If units are correctly randomly assigned, then the groups are “probabilistically similar to each other on the average” (Shadish et al., 2002). If the groups are similar at the outset, then any differences in outcomes can be attributed to the treatment, and an estimate of the size of the effect can also be calculated.
Randomized experiments are often referred to as the gold standard for research (Shadish et al., 2002).

Randomizing subjects to experimental groups theoretically eliminates all bias, or systematic preexisting group differences. Bloom described the attributes of randomization: Each experimental group has the same expected values for all characteristics, observable or not. Randomization of a given sample may produce experimental groups that differ by chance, however. These differences are random errors, not biases. Hence, the absence of bias is a property of the process of randomization, not a feature of its application to a specific sample. The laws of probability ensure that the larger the experimental sample, the smaller preexisting group differences are likely to be. (Bloom, 2006, p. 3)

Without random assignment, the burden is on the researcher to prove that he or she controlled for all of the possible confounding variables that make the treatment group and the control group different (Raudenbush, 2002). Only randomization, which eliminates any confounding factors or systematic differences between treatment and control group members, is seen by many as allowing educational researchers to make strong causal claims about the effects of a particular educational program (Puma, et al., 2009).

Although RCTs are considered the very best, and perhaps the only research designs for making causal inferences, strong quasi-experimental designs, where subjects are not assigned randomly but matched to a comparison group using rigorous statistical methods such as propensity score matching, and using baseline covariates, are also seen as valuable by methodologists (Cook, Scriven, Coryn, & Evergreen, 2010). However, experimental and
quasi-experimental designs, no matter how rigorous the quasi-experimental design, are not seen as equivalent or interchangeable. In a 1996 meta-analysis, Shadish and Ragsdale compared 100 marital and family therapy studies (64 experimental and 36 quasi-experimental) and found that, although the quasi-experimental designs often produce acceptable approximations to experimental designs, particularly when good covariates are used, they consistently yielded less accurate results than the randomized studies. They concluded that, “reliance on results from randomized experiments as the gold standard is still well founded” (Shadish & Ragsdale, 1996, p. 1290).

A simulation study with the same purpose, comparing the results of experimental and quasi-experimental designs, was conducted by Bloom and others at MDRC (Bloom, Michalopoulos, Hill, & Lei, 2002). These researchers used a large sample of data on welfare to work strategies and drew both random and non-random subsamples, then compared the findings from the same analysis conducted on all of the subsamples. Even with sophisticated matching techniques such as propensity score matching, the authors concluded that quasi-experiments produced misleading results and that the biases were often large and unpredictable (Bloom et al., 2002). In talking about the quasi-experimental designs that deployed rigorous matching techniques, the authors stated:

Our results are not encouraging…three of the five in-state comparison groups produced small biases in the short run while two produced large biases. This suggests that an evaluator using in-state comparison groups to assess a…program has a 60 percent chance of getting approximately the right answer and a 40 percent chance of being far off…Adjusting for observed background characteristics did not systematically improve the results. (p. 3-1)
It is important to note that, while RCTs are considered the most rigorous methodology for providing strong, causal evidence for an intervention, reliance on RCTs has been criticized for ignoring the external validity of the findings. In a 2005 Lancet article, Rothwell describes the quandary of clinicians who rely on large-scale, RCTs to identify effective clinical interventions. These studies often are well-constructed in terms of random selection of treatment and control groups, but offer no evidence of external validity or evidence that the findings will generalize to a larger group of patients (Rothwell, 2005).

The Call for Randomized Controlled Trials in Education Research

When the federal agency now known as the U.S. Department of Education was formed in the 1867, part of its mission was to collect statistics and provide exemplary models for the country’s schools. While the design and delivery of education from kindergarten through college has always been primarily a state and local responsibility, federal policymakers who supported the development of the Department of Education felt that a federal agency was the best place to collect and disseminate educational research, as individual states may not see the broader implications of the research being conducted at the state or local level, and therefore not disseminate the findings (U.S. Department of Education, 2011a).

More recently, federal research efforts were led by the creation of the National Institute of Education in the early 1970s, which later became the Office of Educational Research and Improvement (OERI) (Beghetto, 2003). The Education Sciences Reform Act of 2002 replaced the OERI with the Institute of Education Sciences (IES). This change took place soon after 2001 reauthorization of the Elementary and Secondary
Education Act, or No Child Left Behind (NCLB) (Beghetto, 2003; Education Sciences Reform Act of 2002; No Child Left Behind Act of 2001).

Although this federal role in fostering and disseminating educational research has been around since the 19th century, the federal government, by its own admission and others’ admonition, did not live up to the founding ideal of providing exemplary models (Beghetto, 2003; National Research Council, 1999; Whitehurst, 2002a). Leading up to the 2001 reauthorization of NCLB, the National Research Council put forth in its strategic plan for education research and its utilization, the following admission about the state of educational research:

The world of education, unlike defense, health care, or industrial production, does not rest on a strong research base. In no other field are personal experience and ideology so frequently relied on to make policy choices, and in no other field is the research base so inadequate and little used. (National Research Council, 1999, p. 1)

Grover Whitehurst, who was the director of the U.S. Department of Education’s Institute for Education Sciences for several years after NCLB’s passage, illustrated the federal government’s position on the poor quality of educational research that drove NCLB to raise the research bar by comparing the cumulative number of articles using RCTs in education to that of psychology, criminology, and social policy from 1950 through the passage of NCLB in 2002. The cumulative 50-year total number of published research articles based on randomized field trials for education was less than 1,000 for education—at best half that of the other three disciplines examined, which had
between 2,000 and 5,500 published randomized studies over the same period (Whitehurst, 2002b).

Whitehurst served in the first years of the dramatic ratcheting up of standards for educational research under President George W. Bush’s administration with the reauthorization of the Elementary and Secondary Education Act, or NCLB in January 2002 (The No Child Left Behind Act of 2001). NCLB was a fundamental reform of the way public education is delivered in the U.S. Among the many provisions of NCLB is required annual testing in grades 3 through 8 for virtually all students, measures of stepped goals for adequate yearly progress based on student achievement scores, penalties and even reorganization for schools that fail to meet progress goals over time, and the requirement that all tested students be proficient by 2014 (The No Child Left Behind Act of 2001).

The language in the act that is salient to the research proposed for this study is the unprecedented federal government articulation and specification of the kind of research that will be funded, reported, and recommended for school improvement. The act specifies the type of research that will be required for federal funding and dissemination as “scientifically-based research,” defined in the act as “research that involves the application of rigorous, systematic and objective procedures to obtain reliable and valid knowledge relevant to educational activities and programs” (Beghetto, 2003; The No Child Left Behind Act of 2001).

Soon after the passage of NCLB, then IES Director Whitehurst discussed the need for a “gold standard” in education research. He assailed the state of education research
before NCLB as “hit or miss” and full of “folk wisdom” rather than rigorous methods (Whitehurst, 2002a). In a presentation at the 2002 Student Achievement and School Accountability Conference, Whitehurst rank ordered research designs into the quality of the levels of evidence that they provide. RCTs were ranked first on the list for causal research, followed by quasi-experimental designs, then pre-post comparisons, correlational designs, case studies, and lastly anecdotes (Whitehurst, 2002b). Although the NCLB does not outright require RCTs as the only form of research that will be recognized as quality, the gold standard for causal research was clearly set by the act and the ensuing rhetoric from the policymakers and U.S. Department of Education leadership. The most valued research would be conducted as an experimental design executed by a RCT. Whitehurst referred to randomization as “critical” and said without it, differences at outcome could be caused by selection bias (Whitehurst, 2002b). Whitehurst, speaking in his role as Director of the U.S. Department of Education’s Institute for Education Sciences, affirmed in 2003 that, “Randomized trials are the only sure method for determining the effectiveness of education programs and practices.” (Whitehurst, 2003, p. 1)

The U.S. Department of Education’s Institute of Educational Sciences (IES) established the What Works Clearinghouse. The purpose of the WWC is to hold high standards for educational research in order to provide education leaders evidence about whether or not of specific programs and practices have any impact on student achievement (Beghetto, 2003).
The What Works Clearinghouse Procedures and Standards Handbook. In his 2002 testimony before the U.S. House of Representatives Subcommittee on Labor/HHS/Education Appropriations, Grover Whitehurst described the purpose of the WWC:

The What Works Clearinghouse will specify clear and rigorous methodological standards for demonstrations of program effectiveness, and then provide a roster of programs and products, within subject areas, that have met those standards. When sufficient high quality research has accumulated on a particular topic, the clearinghouse will commission synthesis papers that summarize the existing literature and provide guidance on what to expect from the approaches and products that have been subject to research. Over time, the clearinghouse will become the principal source of valid information on effective educational practice. (p. 1)

In order to clarify the process for research to meet the federal government’s rigorous evidence standards, the WWC initially developed and continues to update and provide technical procedural documents such as the Procedures and Standards Handbook and guidance on single case and regression discontinuity designs (Kratochwill et al., 2010; Schochet et al., 2010; U.S. Department of Education, 2008, 2011b). While the four types of studies (RCTs, quasi-experimental designs, single case designs, and regression discontinuity designs) are considered eligible for review and inclusion, the WWC guidance documents indicate an overall preference for RCTs over quasi-experimental designs:
Currently, only well-designed and well-implemented RCTs are considered strong evidence, while quasi-experimental designs (QEDs) with equating may only meet standards with reservations; evidence standards for regression discontinuity and single-case designs are under development. (U.S. Department of Education, 2008, p. 12)

It should be noted that the WWC, when possible, combines or synthesizes the findings of multiple RCTs into summary measures of effectiveness for a particular intervention (U.S. Department of Education, 2013). In 2010 evidence standards for single case and regression discontinuity designs were released (Kratochwill et al., 2010; Schochet et al., 2010).

**Critiques of the What Works Clearinghouse.** The high bar set for educational research to clear scrutiny and be listed as meeting evidence standards by the WWC is critiqued by some as too singular in focus on large, RCTs (Confrey, 2006; Schoenfeld, 2006). Many researchers contend that the relatively low number of studies that have been approved as “meeting evidence standards” by the WWC after almost a decade is detrimental to educational practice when other types of research can add value to the field. Some have argued that the WWC reviews of particular studies have either presented incomplete or misleading information about the effectiveness of the intervention (Confrey, 2006; Greene, 2010; Schoenfeld, 2006).

Confrey (2006) called for a broader discussion among all stakeholders, including educators, methodologists, and content specialists, on what evidence is most helpful and accurate to allow school districts to make important curricular decisions, as well as how
the WWC should present the information so that it is consumable by practitioners who are making curricular purchase decisions that impact student learning. Confrey (2006) took issue with the WWC decision that one large, RCT is always superior to multiple smaller studies or studies with less rigorous designs when evaluating what works in the classroom. Confrey also warned that the average practitioner may confuse the WWC ratings based on whether or a study met the methodological requirements with whether or not a study actually showed that the intervention was effective. Confrey reiterated Campbell and Stanley’s 1963 warning that:

The claims made for the rate and degree of progress which would result from experiment were grandiosely overoptimistic and were accompanied by an unjustified depreciation of non-experimental wisdom….we must increase our time perspective and recognize that continuous, multiple experimentation is more typical of science than once-and-for-all definitive experiments…The experiments we do today will need replication and cross-validation…before they can become theoretically interpreted with confidence. (Campbell & Stanley, pp. 2-3)

It should be noted that Confrey’s critique does not take into account the WWC’s policy of synthesizing evidence of multiple RCTs in determining the extent of evidence for a particular intervention (U.S. Department of Education, 2011; 2013).

The critics of the WWC’s emphasis on RCTs have not changed the fundamental goal of the Clearinghouse, which is to support scientifically-based research on what works in the classroom (U.S. Department of Education, 2011a), so the WWC guidelines continue to have a major impact on the types of educational research being conducted.
Attrition as a Threat to the Validity of Research Findings in Randomized Controlled Trials

For researchers, attrition is almost always present in studies and always unwelcome (Puma et al., 2009; Valentine & McHugh, 2007; Verbeek & Nijman, 1992). At the very least it reduces statistical power and, depending on the type of missing data, can introduce bias or non-random differences between the treatment group and control group. The benefits of randomization at the beginning of the study, including the ability to make causal claims about the effects of an intervention, are lost if the composition of the treatment and control groups changes because of attrition over the course of the study. Attrition may have introduced “non-random differences” between the treatment and control groups (Puma et al., 2009, p. 2). It is seen as “the major threat to longitudinal research” (Miller & Wright, 1995, p. 921).

Shadish and colleagues asserted that attrition is often nonrandom, and that the least acceptable response is to ignore it (Shadish et al., 1998). The magnitude of missing data can be quite large in educational evaluations. Puma and colleagues reviewed impact evaluations funded by the National Center for Educational Evaluation and Regional Assistance and found that outcome data on student achievement are often missing for 10 to 20 percent of the original sample (Puma et al., 2009). McGuigan and colleagues report attrition rates in school-based panel studies of a third or more within three years (McGuigan, Ellickson, Hays, & Bell, 1997). Valentine & McHugh’s 2007 meta-analysis of the effects of attrition on baseline comparability showed up to 30 percent attrition.
**Missing data.** Whether or not attrition has produced nonrandom differences between treatment and control groups is largely based on the assumptions about how the data are missing (Graham & Donaldson, 1993). As described in Chapter 1, a typology of missing data can be used to think about the extent to which the attrition undermines the causal validity of a study’s findings. This missing data typology includes three categories: Missing Completely At Random (MCAR), Missing At Random (MAR), and Missing Not At Random (MNAR). Rubin (1976) cautioned that, “The inescapable conclusion seems to be that when dealing with real data, the practicing statistician should explicitly consider the process that causes missing data far more often than he does” (p. 589).

If it is clear that data are MCAR or even MAR, the loss or attrition is unwanted for statistical power and other reasons, but does not necessarily pose a substantial threat to validity. An analysis of the long-running Michigan Panel Study of Income Dynamics found almost 50 percent attrition over the course of the study for example, but no strong evidence that the attrition had biased the outcomes (Fitzgerald, Gottschalk, & Moffitt, 1997).

The missing data problem in the cases of MCAR and MAR does not necessarily erode the causal validity of the study because the reasons for missing data are unrelated to the effects of the treatment, and there are methods for dealing with MAR scenarios (i.e. multiple imputation and maximum likelihood approaches) (Enders, 2010). However, data missing not at random (MNAR) have the characteristic of attrition being dependent on the outcome variable itself (Enders, 2010; Puma et al., 2009). For example, students with
missing data on the posttest score may be those who would have scored lower than the students who did not drop out of the study. While there is a clear conceptual difference between MAR and MNAR, a practical problem for researchers lies in the fact that it is impossible to know with certainty whether data are MAR or MNAR. For example, there is no way to verify whether or not Hispanic students have more missing outcome data in a reading study than non-Hispanic students because they had poor reading skills or if their missingness was not related to their reading skills because there is no actual outcome data for the students who attrited (Enders, 2010). If a researcher cannot reasonably assume otherwise, then outcome data should be considered missing not at random (Puma et al., 2009).

Another way to categorize and describe attrition relates to the difference in the amount of missing outcome data between the treatment and control groups, or differential attrition. Differential attrition is viewed as the most “troublesome” type of attrition because it can functionally erode an experimental design into a quasi-experimental design and, most importantly, lead to biased results (Yeaton, Wortman, & Langberg, 1983, p. 831). Yeaton and colleagues cited two examples of reanalysis of large-scale, longitudinal studies where it was discovered that differential attrition may have led to incorrect findings (Yeaton et al., 1983). However, the type of missing data (MCAR, MAR, or MNAR) determines the extent of the threat of differential attrition. A recent study modeling the effects of overall and differential attrition in medical trials found that unequal dropout rates did not necessarily imply biased results, and that equal dropout rates did not imply unbiased results (Bell, Kenward, Fairclough, & Horton, 2013).
Indeed, differential attrition that is MNAR is the most troubling for researchers because it poses a serious threat to the validity of the research findings (Shadish et al., 1998; Valentine & McHugh, 2007).

**Methods of addressing missing data.** There are multiple methods for addressing missing data in RCTs, all with the goal of ameliorating the harm caused by the absence of outcome data on a portion of the sample. They range from the simple such as listwise deletion, to the sophisticated such as maximum likelihood estimation, multiple imputation procedures and selection modeling such as the Heckman Model (Fitzgerald, Gottschalk, & Moffitt, 1998; McGuigan et al., 1997; Puma et al., 2009). While Heckman developed his model to address sample selection bias, attrition can be thought of as a particular case of sample selection bias (McGuigan et al., 1997; Miller & Wright, 1995). Heckman’s model first estimates the probability of being observed at posttest, calculates an inverse ratio of that probability. That ratio is then used as a covariate in the regression model (Briggs, 2004; McGuigan et al., 1997). Using a highly-correlated pretest variable as a covariate is another method of addressing missing outcome data (Dong & Lipsey, 2010; Graham & Donaldson, 1993).

Although there are multiple processes for addressing missing data in randomized field trials, there is no consensus among the educational research community about which missing data method should be deployed based on different conditions (Puma et al., 2009). For data that are missing not at random (MNAR), no imputation or correction method has been identified that can reasonably address potential bias. Puma and colleagues’ (2009) simulation of missing data mechanisms for various types of missing
data on both pretests and posttests found that none of the methods commonly used to
impute missing data produced impact estimates with biases of less than 0.05 for data
missing not at random (Puma et al., 2009). This 0.05 bias threshold is the same one set by
the What Works Clearinghouse to make decisions about the validity of the findings of
studies up for review and inclusion into the WWC (U.S. Department of Education, 2008).

The WWC attrition bias protocol. Educational researchers who are
investigating educational interventions in the key topic areas of interest to the WWC
submit their research papers for review in hopes of inclusion in the WWC as “meets
evidence standards” rating. The only studies that can receive that highest rating are
“well-designed and well-implemented randomized controlled trials” (U.S. Department of

The WWC parameters for attrition are based on the premise that the implications
of any attrition bias are related both to the overall rate or magnitude of attrition, as well as
the relationship of the “missingness” to whether a person is in the treatment or control
group, or differential attrition (Puma et al., 2009; U.S. Department of Education, 2008,
2011b). The model the WWC has developed for examining the effects of attrition on a
study are based on a set of assumptions about the relationship between the propensity to
respond and the outcome of interest. They set an “acceptable” level of bias as an effect
size of 0.05 of a standard deviation or less on the outcome variable (U.S. Department of

The principal investigators (now called review team leaders) and certified coders
who are assigned to review a study submitted to the WWC are directed to consider the
types of participants in the study and the likely relationship between attrition and outcome data. This relies on the investigator’s expertise, judgment about the type of participants, and any details described in a report. Investigators may operate on “optimistic” or “pessimistic” assumptions about the relationship between attrition and the outcome of interest. An optimistic assumption may be justified, as illustrated by WWC, if the attrition is in a school-based study of young children where sample loss is likely caused by parent mobility or other factors that are essentially exogenous to the trial. However, if the reviewer thinks that the attrition may be related to the outcome, he or she is directed to use pessimistic assumptions. These assumptions are operationalized if the overall and differential attrition rates place the study in the yellow zone (see Figure 1).

Study authors can make a case that attrition has not undermined the causal validity of the study by presenting evidence, using baseline covariates, that the intervention and comparison groups are still alike, even after attrition (U.S. Department of Education, 2008, 2011b).

**Measurement Error as a Threat to Validity in Randomized Controlled Trials**

In social science research, it is often the case that instruments are developed to measure some construct of human behavior such as mathematical knowledge, depression, or satisfaction. The instruments used to measure these constructs are not without error, which can be systematic or random. Systematic error is error that differs from the true score equally across all study participants, such as when an instrument is not properly calibrated. For example, if a scale weighs two pounds heavier than a person’s actual weight, that extra two pounds will be added to the weight of all study participants.
weighed by that scale. Conversely, random error varies across study participants. The magnitude of measurement error can be quantified and should be reported, as errors in measurement can influence the results of a study (National Research Council, 2002). In educational research, it is nearly always the case that some instrument has been developed to measure student achievement, aptitude, or attitude. This section focuses on the concept of measurement error of instruments designed to measure some construct of human behavior and how that error can be quantified, first with a brief overview of the theory behind the quantification of measurement error, and then with a discussion of that theory operationalized as reliability.

**Classical test theory and the true score.** In classical test theory, the measurement error can be defined as the “discrepancy between an examinee’s observed test score and his or her true score” (Crocker & Algina, 1986, p. 110). The true score is a theoretical construct and can be thought of as the average of all of the observed scores obtained from an individual over an infinite number of repeated testings with the same instrument (Crocker & Algina). Theoretically, it is the person’s score without measurement error. Since no psychological measurement is without error (Kane, 1996), the true score can never be known, only the observed score.

Among the important assumptions about true score and measurement error are that they are uncorrelated and that both are components of the observed score (Haertel, 2006). More precisely, the observed score \( X \) can be thought of as the sum of the true score \( t \) and the error \( e \), as illustrated in Equation 1 (Crocker & Algina, 1986; Kane, 1996; Thissen & Wainer, 2001):
Because of the uncorrelated nature of the components of the observed score, the observed score variance can be separated into two orthogonal components as well (Crocker & Algina, 1986; Kane, 1996; Wainer & Thissen, 2001), as indicated in Equation 2:

\[ \sigma^2_x = \sigma^2_t + \sigma^2_e \] 

This theoretical partitioning of the variance is of an observed score into the true score variance and error variance allows the concepts of reliability and measurement error to be operationalized.

**The reliability of an instrument.** The reliability of an instrument can be defined as its reproducibility or consistency over repeated administrations of the same instrument to the same individuals (Brennan, 2006; Crocker & Algina, 1986). Reliability can also be thought of as the precision of an instrument (Kane, 1996). It is quantified as the ratio of true score (t) variance to observed score (x) variance, as illustrated in Equation 3:

\[ \rho_{xx'} = \frac{\sigma^2_t}{\sigma^2_x} \]
Since the observed score has been defined as the true score plus the error, the observed score variance can be defined as the sum of the true score variance and the observed score variance. Hence, the reliability can be quantified as illustrated in Equation 4:

\[ \rho_{xx} = \frac{\sigma_t^2}{\sigma_t^2 + \sigma_e^2} \quad (4) \]

As evident by Equation 4, if the error variance (\(\sigma_e^2\)) is small, the reliability will be closer to its maximum value of 1. As the error variance increases, the reliability or precision of the instrument decreases.

It is impossible to quantify the exact amount of error in an individual’s observed score, but a range of possible variation in an individual’s score can be calculated. This range of possible variation is called the Standard Error of Measurement, which is derived from the standard deviation of a set of observed scores and the test reliability (Crocker & Algina, 1986).

**Measurement error as a threat to internal validity.** In educational and psychological testing, measurement error is always present (Cook & Campbell, 1979; Kane, 1996). “To a certain extent all psychological measurements are unreliable” (Crocker & Algina, 1986, p. 105). The lack of reliability of an instrument used in a study is one of the threats to the internal validity of the findings because the unreliability of the measure could have caused the treatment effect even if the true effect is zero, or vice versa (Kane, 1996; Shadish et al., 2002). “Because the tolerance for error is a function of
the intended interpretation and use of the measurement procedure, precision is an integral part of validity” (Kane, 1996, p. 355).

**The additive properties of threats to validity.** Shadish, Cook, and Campbell (2002), in their discussion of the threats to internal validity, discussed the “additive and interactive effects of threats to internal validity” (p. 55). They cautioned how the impact of a threat can be added to or depend on the level of another threat. The impact on the net bias depends on both the direction and magnitude of the bias caused by each threat. They “presume that inaccurate causal inferences are more likely the more numerous and powerful are the simultaneously operating validity threats and the more homogeneous their direction” (Shadish et al., 2002, p. 61). Haertel (2006) also articulated the separation between measurement error and other kinds of random error and that they both contribute to the uncertainty of statistical comparisons. Because of this additive property of threats to validity, it is important for any model of the impact of attrition on bias to account for the relative magnitude of actual random error that is affected by good or poor reliability of the instruments used to collect the outcome data.

**Description of Monte Carlo simulations.** Monte Carlo simulations are computer-based methods for approximating values and properties of random variables. Monte Carlo Simulations are in many cases the most viable, and often the only approach to testing assumptions or simulating outcomes because sufficient empirical data are not available. Monte Carlo simulations are often used to test statistical assumptions because the method allows for multiple replications of systematic violations of the assumption. Monte Carlo studies allow for empirical estimation of sampling distribution.
characteristics. Researchers almost never have access to multiple real-world samples with specified characteristics. Monte Carlo procedures allow for researchers to examine the magnitude of the consequences of violating statistical assumptions under varying conditions over repeated samples from known distributions. Monte Carlo methods also allow for determination of the sampling distribution of a statistic when the theoretical distribution is unknown, as in complex statistical analyses involving multivariate statistical techniques (Fan, Felsővályi, Sivo, & Keenan, 2002).

The Monte Carlo method was first used in the 1940s when scientists working on the Manhattan Project needed to model a system that was too complex to solve analytically, then picked up by research labs such as the Rand Corporation after WWII (Thornopoulos, 2013). These simulations are critical when the values and properties of variables of interest are not able to be calculated mathematically or by accessing large samples of real data (Braun & Murdoch, 2007; Brooks, Barcikowski, & Robey, 1999).

As in any standard quantitative research design, Monte Carlo studies involve identifying the population of interest, sampling from that population, collecting the data based on some predefined model or algorithm, then analyzing the results. However, in the special case of Monte Carlo studies, researchers make decisions about how to best simulate the population and then sample from it (e.g., Does a uniform distribution, a standard normal distribution, or another type of distribution best represent the variable of interest?). The model is then run repeatedly (large number of repetitions) drawing from the pseudo-population (Brooks et al, 1999; Thornopoulos, 2013). While there is no set number of repetitions specified for particular uses of Monte Carlo programming, a 1992
study by Robey and Barcikowski looked at Monte Carlo studies of robustness of statistical tests and found that the common practice of carrying out 1,000 iterations was not sufficient, and that substantially larger samples were needed in order to detect violations of robustness. Their study did not recommend a specific number of replications, but provided a range of iterations based on varying assumptions (Robey & Barcikowski, 1992). Current computer capacity and processing speed make it possible to significantly increase the number of iterations used in most Monte Carlo studies.

The difficulties specific to Monte Carlo methods involve computer programming and interpreting the results (Brooks et al., 1999). Monte Carlo methods employ algorithms to generate probability distributions using pseudorandom numbers. The pseudorandom numbers are generated based on a starting number or “seed” that is either selected based on some random starting value such as time of day or specified by the programmer, then placed into a generation formula. Once the probability distributions are generated representing the pseudo-population of interest, the Monte Carlo simulation procedure draws a sample from the distribution or population, calculates a value for the variable of interest, then resamples multiple times in order to get an accurate estimation of the distribution of the statistic of interest. The advantage of the programmer specifying the seed used to generate the pseudorandom numbers is that the process is replicable (Braun & Murdoch, 2007). Using the same seed will generate the same results every time.

In the case of this attrition study, Monte Carlo methods were used by the original researchers and the same methods are used in this paper because it is impossible to use
real data to approximate the effects of attrition on the bias (or variation from the real impact of the treatment on the sample) because the true outcomes of subjects who dropped out of the treatment or control groups before the end of the study are unknown. Pretest data can serve as a proxy for outcome data (Dong & Lipsey, 2010; U.S. Department of Education, 2008, 2011b), and large data sets from RCTs do exist that allow researchers to approximate the impact of attrition on bias of the outcome indicator. However, these data are not available in sufficient numbers to approximate the many combinations of overall and differential attrition. Since all of the possible variations of the impact of attrition on the outcome at follow-up, or posttest score, cannot be computed from actual study data, Monte Carlo methods provide estimations based on approximations of properties of the variables of interest.

**Summary and Conclusions**

The federal government, through the reauthorization of the Elementary and Secondary Education Act in 2002, has become perhaps the biggest change agent in the history of educational research. Citing years of largely poor quality studies with little or no hard data on what works in the classroom, the U.S. Department of Education, through its WWC, moved the bar from fairly low to extremely high for what is considered sound research findings worth considering by policymakers and practitioners. The WWC will rank a study as “meets evidence standards” only if it is a randomized design, and even in the case of these “gold standard” designs, critical reviewers can knock down the ratings based on some key factors—attrition being one of the most important.
Attrition is present when an outcome variable, such as a posttest score, is not available for all participants that were initially randomized into the treatment or control groups. The WWC follows most methodologists’ assertions that attrition is a major threat to the validity of a study’s findings and sets parameters for what is an acceptable level of attrition under liberal and conservative assumptions of how the likelihood of staying in the study is related to a person’s outcome. The WWC gives reviewers some leeway in determining how the data are likely missing. If it cannot be determined that the data are likely missing completely at random or at least conditioned on a variable that is not the outcome of interest (which is almost always impossible to know), the study does not meet evidence standards. This harsh stance is defensible in light of the meta analyses and simulations that indicate no method of imputation or estimation of missing data results in outcome estimates within a range of acceptable bias.

The research and parameters on attrition set by the WWC add value to a topic that has long been seen as critical for methodologists, but largely not addressed systematically or practically with any research-based guidelines until now. The WWC attrition bias modeling and standards arguably move the entire field of educational research forward. One additional topic yet to be addressed specifically by the WWC procedures and standards is that of test reliability. Measurement error is always present and the amount of random error that measurement error inserts varies from instrument to instrument. Even large-scale instruments developed by reputable testing companies for statewide achievement assessment can vary significantly in their reliabilities. For example, the 2009 Fourth Grade Ohio Achievement Assessment battery reliabilities ranged from a .66
for the writing portion to a .89 for the math exam (Ohio Department of Education, 2009). The relative additive impact of test reliability on the bias algorithm used by WWC methodologists can be modeled, adding additional information to the WWC model of threats to validity based on attrition bias.
Chapter Three: Methodology

Research Design

This study utilizes a series of Monte Carlo simulations with a conceptual model (the WWC model) to estimate the bias resulting from varying rates of overall and differential attrition in RCTs. The study first replicates the attrition bias estimates calculated by the U.S. Department of Education’s WWC Procedures and Standards Handbook, Versions 2.0 and 2.1 (U.S. Department of Education, 2008, 2011b), then extends beyond their work by deliberately and more (1) systematically modifying the relationships between the propensity to respond and outcome, (2) adding instrument reliability as a factor, and (3) visually illustrating the impact of varying rates of attrition on bias.

Original study description. The study presented in this paper extends the foundational research first conducted by the U.S. Department of Education and included in their What Works Clearinghouse Procedures and Standards Handbook, Version 2.0 and later in Version 2.1 (U.S. Department of Education, 2008, 2011b). The U.S. Department of Education funded the attrition research in order to provide specific guidance for reviewers of studies seeking to be included in the WWC database. Reviewers must weigh each study using a set of evidence standards, and attrition is one of the standards.

Figure 1, from the WWC Procedures and Standards Handbook, Version 2.1 (2011), roughly illustrates three zones for reviewers to place a study based on the overall and differential attrition reported in the study. The horizontal axis is the percentage of
overall attrition and the vertical axis is the percentage point difference in attrition rates between the treatment and control group. For example, a study with an overall attrition rate of 30 percent and a differential attrition rate of 10 percent would fall into the blue zone.


It is evident from the figure that the WWC considers both overall and differential attrition as contributing to bias, but even a relatively small increase in differential attrition has more of an impact on whether or not a study meets evidence standards than high overall attrition. For example, according to these WWC guidelines, a study can report an overall attrition rate of 40 percent and a differential attrition rate of 2 percent and still be in the green zone, but if the differential attrition rate of that same study were 5 percentage
points higher at 7 percent, the study would fall into the blue zone (U.S. Department of Education, 2011b). If the study has some combination of overall and differential attrition that places it in the blue zone, indicating high levels of potential bias, it can be rated no higher than “meets evidence standards with reservations” and, if reviewers are not provided with evidence that there was baseline equivalence in the treatment and control groups, the study would likely be rejected for inclusion in the WWC. If a study falls in the yellow zone, it is at the discretion of the reviewers to weigh the evidence and decide if the attrition is largely exogenous or endogenous. If they judge that the attrition is likely exogenous or not related to the outcome, as would be the case with first graders in a reading program where the only way they may “drop out” of a classroom-based RCT is if their family moves away, then they can make more optimistic assumptions about the impact of attrition on study outcomes. Conversely, if the reviewers judge that the attrition is most likely endogenous (e.g., high school graduates dropping out of a longitudinal college access study because they didn’t go to college after high school and do not want to report their failure to obtain their goals), they are directed to be more conservative in rating a study acceptable for inclusion in the WWC (U.S. Department of Education, 2008, 2011b).

The attrition protocols were based on a “threshold degree of tolerable bias” set at 0.05 standard deviations of the outcome measure (U.S. Department of Education, 2011b, p. 34). The WWC admits that this 0.05 cutoff is somewhat arbitrary, “There is no right or wrong answer to the amount of bias that can be tolerated” (U.S. Department of Education, 2011b, p. 35). They arrived at this figure by assuming that a study would be
acceptable that reported a 0.25 effect size even though the true effect size might be as low as 0.20. They also report an example of the impact of a study with a bias of 0.05 standard deviations. For a nationally-normed test, a difference of 0.05 standard deviations of the outcome measure results in approximately two percentile point difference for a student at the 50th percentile. If the study reports that the intervention moved a student from the 50th to the 60th percentile (which is a 0.25 effect size), a bias of 0.05 would mean that the student really moved from the 50th to the 58th percentile. For the WWC, that error of 2 percentile points in estimating the effect of the educational intervention would be all they are willing to tolerate (U.S. Department of Education, 2008, 2011b).

While not apparent from the figure, the WWC simulation producing the bias thresholds includes even more complexity than just the relationship between overall and differential attrition. Researchers also include in the model different rates of correlation between a study subject’s propensity to respond or stay in each study condition and the outcome measure. For example, Column 1 in Table 2 indicates a correlation between the propensity to respond and the outcome at follow-up for the treatment group as 0.075 and a correlation between the propensity to respond and the outcome at follow-up for the control group as 0.05, or a slightly weaker correlation than that of the treatment group. Any correlation between the propensity to respond at follow-up and the outcome at follow-up indicates that the data are MNAR. The only way the outcome data could be MCAR is if the correlation between the propensity to respond and the outcome at follow-up was zero for both the treatment and the control groups.
Table 2

What Works Clearinghouse Simulation Results for Bias by Response Rate and Proportion of Outcome Explained by Propensity to Respond (effect size units) (U.S. Department of Education, 2011b, p. 32)

<table>
<thead>
<tr>
<th>P_t</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.900</td>
<td>a_t=0.075</td>
<td>a_t=0.10</td>
<td>a_t=0.15</td>
<td>a_t=0.20</td>
<td>a_t=0.30</td>
<td>a_t=0.50</td>
<td>a_t=1.00</td>
<td>a_t=1.00</td>
</tr>
<tr>
<td>0.890</td>
<td>0.910</td>
<td>0.925</td>
<td>0.935</td>
<td>0.950</td>
<td>0.880</td>
<td>0.790</td>
<td>0.775</td>
<td>0.765</td>
</tr>
<tr>
<td>0.850</td>
<td>0.950</td>
<td>0.925</td>
<td>0.935</td>
<td>0.950</td>
<td>0.925</td>
<td>0.925</td>
<td>0.925</td>
<td>0.925</td>
</tr>
<tr>
<td>0.800</td>
<td>0.900</td>
<td>0.925</td>
<td>0.935</td>
<td>0.950</td>
<td>0.925</td>
<td>0.925</td>
<td>0.925</td>
<td>0.925</td>
</tr>
<tr>
<td>0.790</td>
<td>0.810</td>
<td>0.825</td>
<td>0.835</td>
<td>0.850</td>
<td>0.880</td>
<td>0.900</td>
<td>0.910</td>
<td>0.925</td>
</tr>
<tr>
<td>0.775</td>
<td>0.825</td>
<td>0.835</td>
<td>0.850</td>
<td>0.880</td>
<td>0.900</td>
<td>0.925</td>
<td>0.935</td>
<td>0.950</td>
</tr>
<tr>
<td>0.750</td>
<td>0.800</td>
<td>0.810</td>
<td>0.825</td>
<td>0.850</td>
<td>0.880</td>
<td>0.900</td>
<td>0.910</td>
<td>0.925</td>
</tr>
<tr>
<td>0.700</td>
<td>0.700</td>
<td>0.710</td>
<td>0.725</td>
<td>0.735</td>
<td>0.750</td>
<td>0.780</td>
<td>0.800</td>
<td>0.825</td>
</tr>
<tr>
<td>0.690</td>
<td>0.710</td>
<td>0.725</td>
<td>0.735</td>
<td>0.750</td>
<td>0.780</td>
<td>0.800</td>
<td>0.825</td>
<td>0.850</td>
</tr>
<tr>
<td>0.675</td>
<td>0.725</td>
<td>0.735</td>
<td>0.750</td>
<td>0.780</td>
<td>0.800</td>
<td>0.825</td>
<td>0.850</td>
<td>0.880</td>
</tr>
<tr>
<td>0.665</td>
<td>0.735</td>
<td>0.750</td>
<td>0.780</td>
<td>0.800</td>
<td>0.825</td>
<td>0.850</td>
<td>0.880</td>
<td>0.900</td>
</tr>
<tr>
<td>0.650</td>
<td>0.750</td>
<td>0.780</td>
<td>0.800</td>
<td>0.825</td>
<td>0.850</td>
<td>0.880</td>
<td>0.900</td>
<td>0.925</td>
</tr>
<tr>
<td>0.600</td>
<td>0.660</td>
<td>0.710</td>
<td>0.735</td>
<td>0.750</td>
<td>0.780</td>
<td>0.800</td>
<td>0.825</td>
<td>0.850</td>
</tr>
<tr>
<td>0.590</td>
<td>0.610</td>
<td>0.635</td>
<td>0.675</td>
<td>0.710</td>
<td>0.735</td>
<td>0.760</td>
<td>0.785</td>
<td>0.810</td>
</tr>
<tr>
<td>0.575</td>
<td>0.625</td>
<td>0.675</td>
<td>0.710</td>
<td>0.750</td>
<td>0.790</td>
<td>0.825</td>
<td>0.860</td>
<td>0.890</td>
</tr>
<tr>
<td>0.565</td>
<td>0.635</td>
<td>0.675</td>
<td>0.710</td>
<td>0.750</td>
<td>0.790</td>
<td>0.825</td>
<td>0.860</td>
<td>0.890</td>
</tr>
<tr>
<td>0.550</td>
<td>0.650</td>
<td>0.675</td>
<td>0.710</td>
<td>0.750</td>
<td>0.790</td>
<td>0.825</td>
<td>0.860</td>
<td>0.890</td>
</tr>
</tbody>
</table>

The modeling that produced Figure 1 includes a range of eight assumptions about the relationship between the propensity to respond and outcome. Table 2 illustrates all of the components of the WWC attrition model, including the response rate (P_t and P_c, subscripts indicate treatment and control groups) in the table, plus the eight categories illustrating different magnitudes of the relationship between the propensity to respond and outcome at follow up (a_t and a_c). A total of 160 bias estimates, in effect size units, are illustrated in the table. The bias estimates, in effect size units, are defined as “the
difference in mean outcomes between treatment and control respondents” (U.S. Department of Education, 2011b, p. 32). Since the simulation that developed these bias estimates differentiated between treatment and control participants’ response at outcome only on the basis of attrition and the magnitude of the correlation between the propensity to respond and outcome (and not on actual differences in the outcome measure), any difference in mean outcomes between the treatment and control groups is bias, as no true difference in outcomes is modeled.

The table includes overall response rates ranging from 90 percent to 60 percent, which can be defined as overall attrition rates of 10 percent through 40 percent. The first five rows of the table, for example, are at the 90 percent response rate or 10 percent overall attrition rate level, but the modeling illustrates five scenarios of differential attrition at that level. In the row 1 ($P_t = 0.90$ and $P_c = 0.90$) there is no differential attrition. In row 5 ($P_t = 0.85$ and $P_c = 0.95$), the overall attrition rate is at 0.90 and the differential attrition rate is at 0.10.

The columns indicate the eight scenarios of the relationship between the propensity to respond and outcome. They vary from a relatively low relationship between the propensity to respond and the outcome measure in Column 1 ($\alpha_t = .05$ and $\alpha_c = 0.075$) and a perfect relationship between the propensity to respond and the outcome in Columns 7 and 8 ($\alpha_t = 1$ and $\alpha_c = 1$ or -1). These are correlational statistics and are defined as how much the outcome is correlated with a respondent’s likelihood, or propensity to respond to the outcome measure (not drop out of the study). For example, high school graduates who were enrolled in a college access program and that did not go
to college immediately after graduation may be less likely than the control group to respond to a follow-up survey asking about their college enrollment status because they do not want to report that they did not follow through on their plans. In this case, the attrition should be viewed as largely endogenous and the $\alpha_t$ value would be large, and larger than the $\alpha_c$ value. Conversely, if attrition can be viewed as completely exogenous, as in a weather-related school closing causing some students to miss the posttest, then the values of both $\alpha_t$ and $\alpha_c$ would be zero, indicating no relationship between the outcome at follow up and the propensity to respond.

Monte Carlo studies require researchers to make choices about the structure of the simulation that enables close modeling of the real system. They must decide on the equations and algorithms as well as the correct probability distributions to use to simulate the variables. When possible, this matching of the simulation to real systems is accomplished by using existing empirical data that are available, and often times the choices are dependent upon the “best judgment of one or more experts” (Thornopoulos, 2013, p. 4). The WWC Procedures and Standards Handbook offers no explanation for how they determined the eight sets of relationships between the propensity to respond and outcome (the $\alpha_t$ and $\alpha_c$ values). Indeed, the need for Monte Carlo simulation methods to research attrition is driven by the fact that there is insufficient evidence from enough existing datasets to look at the impact of overall and differential attrition interacting with the relationship of the likelihood of responding and the outcome of the study. It is impossible to truly quantify the relationship between the propensity to respond and the outcome because the outcome for nonrespondents is almost always unknowable.
However, Mathematica Policy Research, the research group conducting the attrition study on behalf of the U.S. Department of Education, attempted to ground truth or test their model by using pretest data from a large RCT as a proxy for posttest or outcome data. Their goal was to see if their assumptions about the relationship between the propensity to respond and outcome were tenable.

In order to test the accuracy of their model, the developers used a Mathematica Policy Research evaluation of an educational technology program that had four separate interventions in four grade levels (U.S. Department of Education, 2008, 2011b). Across the four outcome measures, the average difference (in effect size units) between respondents and nonrespondents was 0.40, and the difference between the treatment and control groups in the respondent-nonrespondent difference was an average of 0.10 (this was in Column 1 where the correlation between the propensity to respond and the outcome for both treatment and control groups was lowest \( \alpha_t = 0.075 \) and \( \alpha_c = 0.05 \)). Applying that empirical finding to the simulation data, the WWC researchers used the 90 percent response level and no differential attrition (row 1 in Table 2), and calculated the overall difference between respondents and nonrespondents as well as the difference between the treatment group and the control group in the difference between respondents and nonrespondents. Table 3 illustrates the findings of that component of the WWC simulation study (U.S. Department of Education, 2008).
Table 3

What Works Clearinghouse Simulation Results Illustrating Overall Differences between Respondents and Nonrespondents and the Difference in that Difference between the treatment and Control Groups in the Case of 90% Response and No Differential Attrition (U.S. Department of Education, 2011b, p. 33)

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
<th>(7)</th>
<th>(8)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha_t$=0.075</td>
<td>$\alpha_c$=0.10</td>
<td>$\alpha_t$=0.15</td>
<td>$\alpha_t$=0.20</td>
<td>$\alpha_t$=0.30</td>
<td>$\alpha_t$=0.50</td>
<td>$\alpha_t$=1.00</td>
<td>$\alpha_c$=1.00</td>
<td></td>
</tr>
<tr>
<td>$\alpha_t$=0.05</td>
<td>$\alpha_c$=0.05</td>
<td>$\alpha_c$=0.05</td>
<td>$\alpha_c$=0.15</td>
<td>$\alpha_c$=0.20</td>
<td>$\alpha_c$=0.20</td>
<td>$\alpha_c$=1.00</td>
<td>$\alpha_c$=-1.00</td>
<td></td>
</tr>
</tbody>
</table>

| Difference between all respondents and nonrespondents | 0.49 | 0.52 | 0.60 | 0.81 | 0.97 | 1.12 | 1.95 | 0.00 |
| Difference between the treatment and control groups in the difference between respondents and nonrespondents | 0.10 | 0.18 | 0.32 | 0.12 | 0.20 | 0.50 | 0.00 | 3.90 |

The closest match of the simulated data to the Mathematica Policy Research study data was in column 1, both in the difference between ALL respondents and nonrespondents (0.49 effect size units) and the difference between the treatment and control groups in that overall difference (0.10 effect size units). This is the scenario with a very high response rate (90 percent), no differential response rate, and little correlation between the propensity to respond and the outcome. The researchers concluded that assuming little correlation between the propensity to respond and outcome in the Mathematica Policy Research study was valid because it was a classroom-based study where students had no choice about whether or not to be in a classroom that used a
particular educational technology that was the treatment (U.S. Department of Education, 2008). In sum, the values for at least one cell in Table 3 yielded what the WWC researchers concluded was reasonable compared to a real data set where attrition rates were known and outcomes for nonrespondents could be approximated using pretest data.

In order to generate the values in Table 2, the WWC researchers used a Monte Carlo simulation written in the R programming language (The R Development Core Team, 2008) to generate probability distributions for the following variables that are defined later in this chapter: 1) propensity to respond (standard normal distribution); 2) quantile function (inverse of the cumulative distribution function); and 3) factors unrelated to attrition, or the error term (standard normal distribution). The inputs to the model were: 1) four different sets of overall response rates, with each of the four sets containing five variations of differential response rates between the treatment and control groups (the values for \( P_t \) and \( P_c \) in Table 2); and 2) eight different scenarios of the relationship between the propensity to respond and the outcome (the values for \( \alpha_t \) and \( \alpha_c \) in Table 2).

Second study description. The first research to replicate and extend the WWC attrition bias model was that of Dong and Lipsey (2010). They used a macro program written in SAS 9.2 to first replicate the WWC model of attrition bias, then extend the simulation to address the following additional research questions: 1) Does a well-chosen covariate reduce bias caused by attrition? 2) What is the nature and influence of attrition in cluster RCTs (Dong & Lipsey, 2010). The second research question was articulated by the researchers who developed the What Works Clearinghouse Procedures and Standards

To first replicate the WWC simulation and then address their additional research questions, Dong and Lipsey (2010) used a Cholesky decomposition (a method of creating correlated data) to the variance-covariance matrix (Fan et al., 2002) to create 5,000 correlated random variables (2,500 for the treatment group and 2,500 for the control group). They needed these correlated random variables for the examination of the impact of the pretest as a covariate. From this master dataset of 5,000 correlated random variables, they generated 1,000 master datasets for each parameter combination of the respondent sample.

With the generated data sets, Dong and Lipsey first replicated the WWC attrition estimates (the model with no covariate), then in order to assess the impact of the covariate on the estimated bias due to attrition, they compared the results of the modeling without the random variable that serves as a pretest (the original WWC results) to the model with the pretest.

To address their second research question, Dong and Lipsey (2010) modeled attrition bias that could be present in a cluster RCT where outcome data could be missing at the individual level (level 1), the cluster level (level 2), or both. In order to examine this, they use the SAS 9.2 macro program to generate a two-level dataset by specifying intraclass correlations (the correlations among individuals within the same cluster). In order to assess the impact of a covariate (pretest data) on attrition bias, they again deployed Cholesky decompositions to create correlated random variables at level 1 and
level 2. They create 100 clusters with 50 individuals per cluster (50 for the treatment group and 50 for the control group). Then, for each parameter combination (combination of attrition rates and relationship of the propensity to respond and outcome), they generated 1,000 datasets from the correlated, multilevel data they generated. They then examined the nature of bias for the clusters with attrition at both level 1 and level 2.

Dong and Lipsey’s overall conclusion was that, both for randomized and cluster RCTs, the bias introduced by the differential response rate between the treatment and control groups is reduced if baseline covariates that are correlated with the outcome variable and propensity to respond are included in the data analysis (Dong & Lipsey, 2010). An example of such a covariate is a pretest score. It can be assumed to correlate with the propensity to respond as well as a proxy for the posttest score or outcome. For example, students with lower pretest scores may feel less capable of completing a voluntary intervention and drop out of the study at greater rates than those scoring higher at pretest.

Although the inclusion of a correlated covariate may help reduce bias in some attrition scenarios driven by differential attrition, Dong and Lipsey also concluded that the inclusion of a correlated covariate may not have much impact on the overall bias if the larger source of that bias is introduced by overall rather than differential attrition. For cluster RCTs, the impact of bias due to attrition at both levels is more complex, and attrition at level two may have less impact on bias than attrition at level 1, especially if the intraclass correlation for the group is relatively small (Dong & Lipsey, 2010).
Because their study purpose was to first replicate, then elaborate on the model developed for the WWC and because their work was primarily for an academic audience, Dong and Lipsey not only extended the research on attrition bias, but also helped to specify the procedures and assumptions used in the development of the WWC attrition models. The WWC Procedures and Standards Handbook was written not as an academic paper, but with the purpose of guiding researchers who seek to get their work reviewed and included in the WWC, as well as for WWC reviewers. Consequently, the WWC simulation provided little detail on how they operationalized the model. The Dong and Lipsey study provided detailed explanations of the equations used in the simulations and resulting in the estimates of bias, which they were able to replicate.

For example, the WWC researchers described bias in the model as being generated by both the differential response rate in the treatment and control groups and the proportion of variation of the outcome variable explained by the propensity to respond. Also, the WWC model provided values for the proportion of the outcome that can be explained by the propensity to respond and describes them in a footnote as the regression $R^2$. Dong and Lipsey’s paper clarified that it is the square root of those given values are utilized in the OLS regression equations that derive the bias estimates (Dong & Lipsey, 2010; U.S. Department of Education, 2008).

**The Statistical Model**

This section lists the specific variables that will be included in the simulation and the next section defines the operational definition of each variable. All of the data for this study were generated using the software package R (The R Development Core Team,
In order to first replicate the WWC attrition bias effect size estimates and modify the parameters to address the specific research questions in this study, the following variables were simulated or specified in the model:

1) **Attrition**
2) **Bias**
3) **Attrition Bias**
4) **Propensity to Respond**
5) **Proportion of Individuals who Respond**
6) **Outcome at Follow-up**
7) **Factors Unrelated to Attrition**
8) **Proportion of Variation in the Outcome that can be Explained by the Propensity to Respond**
9) **Error Introduced by Reliability of the Instrument**

**Operational Definition of the Variables**

**Attrition.** Attrition in RCTs is the loss of participants or units after random assignment has taken place. Attrition can occur before, during, or after treatment. Attrition occurs when the participant drops out before data on the outcome measure can be collected, during any point after random assignment (Schafer & Graham, 2002).

**Bias.** Bias, in a statistical framework, is a systematic error in an estimate or inference (Shadish et al., 2002) or difference between two parameters that should be equal (Camilli, 2006). For purposes of these studies, the WWC defined bias as the
deviation from the true impact, and thereby a threat to causal validity, for the analysis sample (U.S. Department of Education, 2008, 2011b).

**Attrition bias, \( y_t - y_c \).** Dong and Lipsey (2010) defined attrition bias as “the difference between the empirical impact point estimate using the respondent sample and the real impact on the whole sample” p. 4. They also parsed the overall attrition bias into two components: (a) Part 1 bias- the bias introduced by the systematic difference between the original whole sample and the respondent sample. This is the overall response/attrition rate for the study and is a threat to external validity, as the respondent sample may not be representative of the original random sample from the population of interest, thereby reducing the value of randomization in the RCT; (b) Part 2 bias- the bias introduced by any differential response rate between the treatment and control groups, which is a threat to internal validity (Clarke, 2005; Rosenbaum, 1986, 2002) since unobserved differences between the treatment and control groups may impact the outcome as much or more than the treatment, so any causal inference is compromised.

In the WWC study and in this study, bias is the difference between \( y_t \) and \( y_c \) among respondents, which is generated by response rates (\( \rho_t \) and \( \rho_c \)) and the relationship between the propensity to respond and the outcome (\( \alpha_t \) and \( \alpha_c \)) for the treatment and control groups (U.S. Department of Education, 2008).

**Propensity to respond, \( z \).** In the WWC model, the propensity to respond (\( z \)) is a latent variable (unobserved) with a standard normal distribution (N[0,1]). If an individual’s propensity to respond exceeds a threshold derived from the response rate (\( \rho \)) using the quantile function, then the individual responds. In other words, an individual
responds at posttest (does not drop out of the study) if that person’s propensity to respond \( (z) \) is greater than the value that corresponds to a specific percentile of the \( z \) distribution, given the response rate. In the WWC study and in this study, the propensity to respond is generated through simulation using the standard normal distribution and the quantile function.

**Proportion of individuals who respond, \( (p) \).** This is the response rate and can be different for the treatment and control groups. In the WWC study and in this study, the researchers input varying overall and differential response rates into the model.

**Outcome at follow-up, \( (y) \).** The WWC researchers define this as the “key quantity of interest” (p. 29). They model it as the sum of two unobserved quantities: (a) a factor unrelated to attrition, or the random error; and (b) the propensity to respond. This model assumes no effect of the treatment on the outcome and includes no covariates. It is operationalized as the linear equation:

\[
y = \alpha \cdot z + (1-\alpha) \cdot u
\]

**Factors unrelated to attrition, \( (u) \).** In the WWC model, this is an unobserved variable (simulated) that represents all factors of the outcome \( (y) \) unrelated to attrition. Since the WWC model assumes that there is no effect of the treatment on the outcome, by extension, this quantity would include the random error in the measurement of the outcome \( (y) \). It is simulated from a standard normal distribution (N[0,1]).

**Proportion of the variation in the outcome that is explained by the propensity to respond, \( (\alpha) \).** This is the relationship between the propensity to respond and outcome in the WWC simulation. It can be interpreted similarly to a regression \( R^2 \). The larger the
α value, the more a participant’s response (or possible response) is related to his or her score on the outcome variable. This can also be conceptualized as the extent that the outcome data are MNAR. For example, in a weight loss study, persons who have not lost weight may be less likely to stay in the treatment or respond to the post-survey because they feel like they failed in their goals; persons who have lost the most weight may be eager to respond with their achievements. In this case, α would be large. If α = 1, the entire outcome is explained by the propensity to respond. If α = 0, none of the outcome is explained by the propensity to respond, as illustrated here:

\[ y = \alpha \cdot z + (1-\alpha) \cdot u = u \]  

Moreover, the proportion of individuals responding on the outcome variable may differ by treatment status, as illustrated here:

\[ Y_t = \alpha_t \cdot z_t + (1-\alpha_t) \cdot u_t \]  
\[ Y_c = \alpha_c \cdot z_c + (1-\alpha_c) \cdot u_c \]

If α is equal in both groups, then equal rates of attrition in both groups do not harm the internal validity of the study, as the same kind of participants are assumed to drop out of both groups. However, external validity may still be compromised, as the participants who drop out may be different (in both groups equally) from those who do not drop out (U.S. Department of Education, 2008).
Simulating differences in the reliability of the instrument across studies (ur).

This concept introduced in the current research and represents the variability across different studies in the measurement error or lack of precision of the outcome instrument, which is a component of the random term in participant’s outcome score at follow up, or \( y \). It is commonly referred to as the reliability of the instrument and is quantified by a reliability coefficient. In classical test theory, the measurement error can be defined as the “discrepancy between an examinee’s observed test score and his or her true score” (Crocker & Algina, 1986, p. 110). The true score is a theoretical construct and can be thought of as the average of all of the observed scores obtained from an individual over an infinite number of repeated testings with the same instrument (Crocker & Algina, 1986). Theoretically, it is the person’s outcome without random measurement error.

Outcome at follow up \( y \) for each individual in the WWC attrition study is defined as a factor unrelated to attrition. Since the WWC model assumes no actual difference in the outcome measure (the true score) between the treatment and control groups (U.S. Department of Education, 2008, 2011b) the factor unrelated to attrition in the \( y \) term can be defined as the random error of the outcome.

The WWC attrition modelers used a standard normal distribution to model the effects of that random error term on the participant’s outcome at follow-up. There is no accounting in the WWC attrition model to simulate the differential relative impact of the different reliabilities of the outcome instruments across studies. This current study modeled the effects of differences in reliability of the instrument across studies and its potential impact on bias by simulating relatively wider variations in the random error that
may be introduced in some studies that utilize outcome instruments with relatively low reliability.

The reliability coefficient in the classical true score model is described as the proportion of observed score variance that is attributable to the variance in test takers’ true scores (Crocker & Algina, 1986). It is a statistic ranging from zero to one, and is interpreted as the expected correlation between two replications of the same test to the same subjects. Since the true score is a theoretical concept and cannot be known precisely, the reliability coefficient needs to be estimated. Reliability is estimated by a number of methods, including those that involve administering parallel forms of the same test and those that divide items from a single administration of a test into two parts and estimate the internal consistency of the instrument (Crocker & Algina, 1986; Haertel, 2006).

Kane, in his 1996 article, “The Precision of Measurements,” discussed the degree of measurement error that can/should be tolerated as context-specific. As stated by Kane (1996):

Tolerance is closely connected to the concept of validity because the tolerance for error is determined by the impact of errors of measurement on interpretations and decisions…To ensure that random errors of measurement are not undermining the intended outcomes, it is necessary to show that the errors are small compared with the tolerance for errors implied by the proposed application. Therefore, an evaluation of the magnitude of errors relative to the tolerance for error is an integral part of any validation effort. The assertion of classical test theory that
reliability is necessary but not sufficient for validity is a special case of this general conclusion. (pp. 359-360)

Shadish, Cook, and Campbell (2002) described the "additive and interactive effects of threats to internal validity" p. 55. They cautioned how the impact of a threat to validity can be added to or depend on the level of another threat. No instrument is without measurement error. It is impossible, however, to quantify the exact amount of error in an individual’s observed score, but a range of possible variation in an individual’s score can be calculated. This range of possible variation is called the Standard Error of Measurement, which is derived from the standard deviation of a set of observed scores and the test reliability (Crocker & Algina, 1986).

In order to simulate the differential impact on the outcome at follow-up ($y$) across studies based on the differences in reliabilities of the instruments used, this study increased the standard deviation of the distribution used to insert the ($u$) term, or the factor unrelated to attrition.

**The Simulation**

In order to first replicate the WWC simulation, then extend beyond it to address the additional research questions regarding sensitivity to variations in the scenarios and the variations in the reliability of the outcome instruments across studies, this study utilized the R programming language to generate the following variables, functions, distributions:

1) $z_t$ an individual’s propensity to respond in the treatment group, or the extent that the outcome data are MNAR. The simulation drew $z_t$ values from an R-
generated standard normal distribution with a mean of 0 and a standard
deviation of 1 (N[0,1]), as this is the assumed distribution of \( z \) based on the
WWC model. This was accomplished by deploying the \textit{rnorm} function in R.

2) \( z_c \) - an individual’s propensity to respond in the control group, or the extent
that the outcome data for the control group are MNAR. The simulation drew
\( z_c \) values from a separate, R-generated standard normal distribution, N(0,1)

3) \( Q \) - the quantile function, or the inverse cumulative distribution function based
on a given response rate, or \( \rho \). For each value of \( z \) drawn in the simulation,
the quantile function (the \textit{qnorm} function in R) was deployed to see if the \( z \)
value drawn exceeds a particular percentile of the previously-generated \( z \)
distribution, based on the given, actual response rate, or \( \rho \) (\textit{qnorm}[\rho]). If the \( z \)
value drawn in that particular loop of the program exceeds the percentile of
the \( z \) distribution at a given \( \rho \), then that draw is classified as a respondent and
an outcome value for \( y \) was calculated for that respondent. This was
simulated for both the treatment and the control groups, as illustrated in
equations 1 and 2 later in this section. It was simulated for every given
combination of response rate and relationship of response to outcome.

4) \( u_t \) - factors unrelated to attrition for the treatment group. The simulation drew
\( u_t \) values from an R-generated standard normal distribution with a mean of 0
and a standard deviation of 1 (N[0,1]), as this is the assumed distribution of \( u \)
based on the WWC model. This was accomplished by deploying the \textit{rnorm}
function in R.
5) $u_c$ - factors unrelated to attrition for the control group. The simulation drew $u_c$ values from an R-generated standard normal distribution with a mean of 0 and a standard deviation of 1 ($N[0,1]$), as this is the assumed distribution of $u$ based on the WWC model. This was accomplished by deploying the \texttt{rnorm} function in R.

6) $u_{t_r}$ - for the research question addressing the potential for variation in reliability of instruments across studies that introduces more or less additional error introduced by the lack of reliability of the instrument for the treatment group. The simulation will draw $u_t$ values from an R-generated normal distribution with a mean of 0 and a standard deviation of 1.2. This was accomplished by deploying the \texttt{rnorm} function in R.

7) $u_{r_c}$ - additional error introduced by the lack of reliability of the instrument for the control group. The simulation drew $u_t$ values from an R-generated normal distribution with a mean of 0 and a standard deviation of 1.2. This was accomplished by deploying the \texttt{rnorm} function in R.

In addition to the variables generated through the above distributions and functions, specific values for response rate (the $\rho_t$ and $\rho_c$ value combinations in Table 2) and the relationship between the propensity to respond and outcome ($\alpha_t$ and $\alpha_c$) were first taken directly from the WWC simulation in order to replicate the specific findings, then augmented with more systematic increases and differences between treatment and control groups and relationship between the propensity to respond and outcome.
Every replication (n=50,000) inserted the draws from the distributions of \( z, u \) (or \( ur \)) for each \( \rho \) and \( \alpha \) combination into the following formulas:

\[
\text{If } z_t > Qz (1-\rho_t), \text{ then } y_t = \alpha_t \ast z_t + (1-\alpha_t) \ast u_t \\
\text{If } z_c > Qz (1-\rho_c), \text{ then } y_c = \alpha_c \ast z_c + (1-\alpha_c) \ast u_c
\]

(9) (10)

For the separate simulation modeling variations across studies in the magnitude of the random error based on the reliability of the outcome instrument, the same formulas were used with different notations:

\[
\text{If } z_t > Qz (1-\rho_t), \text{ then } y_t = (\alpha_t \ast z_t + (1-\alpha_t) \ast ur_t) \\
\text{If } z_c > Qz (1-\rho_c), \text{ then } y_c = (\alpha_c \ast z_c + (1-\alpha_c) \ast ur_c)
\]

(11) (12)

Once the \( y \) values for the treatment and control groups were generated, the mean of \( y_t \) and the mean of \( y_c \) was calculated and the bias is estimated using the following formula:

\[
\text{Bias} = \text{mean}y_t - \text{mean}y_c
\]

(13)

All cells were populated based on 50,000 replications of the program. This high number of replications in a Monte Carlo simulation provides more stable estimates than
those generated by the 5,000 replications used by the WWC researchers (Robey & Barcikowski, 1992).

**Simulation Scenarios**

The simulations deployed in this study produced multiple sets of bias estimates that address or inform each research question. Each research question was addressed by a replication or modification of the original Mathematica Policy Research model. The specific modifications of the model to address the research questions are described here:

**Problem 1: Do the amount and type of attrition, under varying assumptions of how much subjects’ likelihood of responding is related to their outcome, impact randomized controlled studies by contributing to systematic bias as in the original study?** This research question requires a replication of the original WWC study and outcomes. The researchers who developed the attrition bias simulation did not provide any of their R programming syntax, so was important to first replicate the WWC results in order to ensure that this study models the attrition bias estimates in the same way as the WWC researchers. Otherwise, any extension of the model to address additional research questions would not be based on the same fundamental assumptions of the relationship between attrition and bias.

More precisely, using the WWC simulation’s inputs and formulas, this in Simulation Scenario 1, reproduced the results from Table 2 that are included in the WWC Procedures and Standards Handbook, versions 2.0 and 2.1 (U.S. Department of Education, 2008, 2011). The ability to replicate the WWC researchers’ simulation is the
first research problem because is the foundational research supporting the rest of this study.

Simulation scenario 1- addressing problem 1. This first simulation replicated the WWC Monte Carlo study with no modifications to the model inputs. The simulation produced a matrix of bias estimates (replicating Table 2) at varying levels of overall and differential attrition as well as varying levels of the magnitude and relative magnitude of outcome data MNAR (or relationship of the propensity to respond to outcome). The algorithm for the WWC model first simulated the distributions for the variables in equations 7 and 8, performed 50,000 draws from those variables and solved for \( Y_t \) and \( Y_c \), and finally estimated the bias by subtracting the mean outcome at follow-up for the treatment group from the mean outcome at follow-up for the control group (Bias = mean\( Y_t \) – mean\( Y_c \)). The formulas are based on the formulas provided in the WWC Procedures and Standards Handbook, Version 2.1 (U.S. Department of Education, 2011).

\[
Y_t = \alpha_t * z_t + (1-\alpha_t) * u_t \quad (7)
\]
\[
Y_c = \alpha_c * z_c + (1-\alpha_c) * u_c \quad (8)
\]

The vectors of \( \alpha_t \) (propensity to respond in the treatment group) and \( \alpha_c \) (propensity to respond in the control group) are:

\( \alpha_t \) - (.075, 0.1, 0.15, 0.2, 0.3, 0.5, 1, 1)

\( \alpha_c \) - (0.05, 0.05, 0.05, 0.15, 0.2, 0.2, 1, -1)
The vectors $\rho_t$ (response rate of treatment group) and $\rho_c$ (response rate of control group) are:

$\rho_t = (0.9, 0.89, 0.875, 0.865, 0.85, 0.8, 0.79, 0.775, 0.765, 0.75, 0.7, 0.69, 0.675, 0.665, 0.65, 0.6, 0.59, 0.575, 0.565, 0.55)$

$\rho_c = (0.9, 0.91, 0.925, 0.935, 0.95, 0.8, 0.81, 0.825, 0.835, 0.85, 0.7, 0.71, 0.725, 0.735, 0.75, 0.6, 0.61, 0.625, 0.635, 0.65)$

**Problem 2: How do changes in the assumptions about the correlation between the propensity to respond and outcome and the magnitude of differential attrition affect bias in randomized controlled trials?** The WWC researchers used eight scenarios of relationship between the propensity to respond (getting an outcome measure on a participant) and the outcome for that respondent. This relationship is notated with $\alpha_t$ and $\alpha_c$. This study more systematically steppe the values for $\alpha_t$ and $\alpha_c$ as well as $P_t$ and $P_c$ in order to more precisely and thoroughly model the sensitivity of the bias estimates based on different rates of overall and differential attrition as well as the correlation between the outcome value and whether or not a participant drops out of the study.

**Simulation scenarios 2 through 6- addressing problem 2.** Scenario 2 simulated the counterfactual, or bias estimates in the absence of any relationship of the propensity to respond to outcome (no data MNAR). The scenario first simulated the complete counterfactual, with no attrition and no relationship between the propensity to respond and outcome, then added the same magnitude of overall and differential attrition as the original WWC simulation. The simulation model remained the same. The only change in
the model was the input values for $\alpha_t, \alpha_c, \rho_t,$ and $\rho_c$. The vectors of $\alpha_t$ (propensity to respond in the treatment group) and $\alpha_c$ (propensity to respond in the control group) are:

$\alpha_t$ - (0)

$\alpha_c$ - (0)

The vectors $\rho_t$ (response rate of treatment group) and $\rho_c$ (response rate of control group) are:

$\rho_t$ - (1, .9, .89, .875, .865, .85, .8, .79, .775, .765, .75, .7, .69, .675, .665, .65, .6, .59, .575, .565, .55)

$\rho_c$ - (1, .9, .91, .925, .935, .95, .8, .81, .825, .835, .85, .7, .71, .725, .735, .75, .61, .625, .635, .65)

Scenario 3 more systematically stepped the $\alpha_t$ and $\alpha_c$ values which are the correlation between the outcome value and whether or not a participant drops out of the study (the magnitude of the data that are MNAR). Specifically, more increments were added to the treatment group propensity to respond, or $\alpha_t$, and the control group propensity to respond, or $\alpha_c$, was increased in increments of .05 for each of the values of the treatment group propensity to respond ($\alpha_t$) without equaling or exceeding the $\alpha_t$ value, starting with the $\alpha_t$ value of .075 and ending with the $\alpha_t$ value of .20. There was no research-based rationale for the incremental steps in the relationship between the outcome value and whether or not a participant drops out of the study. The stepped values were arbitrary, but uniform.
(The simulation model remained the same. The only change in the model was the input values for $\alpha_t$ and $\alpha_c$. The vectors of $\alpha_t$ (propensity to respond in the treatment group) and $\alpha_c$ (propensity to respond in the control group) are:

$\alpha_t = (.075, .075, .085, .085, .095, .095, 10, 10, 15, 15, 20, 20, 20)$

$\alpha_c = (.05, .10, .05, .10, .05, .10, .05, .10, .05, .10, .05, .15)$

The vectors $\rho_t$ (response rate of treatment group) and $\rho_c$ (response rate of control group) are (same as original WWC model):

$\rho_t = (.9, .89, .875, .865, .85, .8, .79, .775, .765, .75, .7, .69, .675, .665, .65, .6, .59, .575, .565, .55)$

$\rho_c = (.9, .91, .925, .935, .95, .8, .81, .825, .835, .85, .7, .71, .725, .735, .75, .6, .61, .625, .635, .65)$

Scenario 4 continued (from Scenario 3) to extend the systematic increase in the $\alpha_t$ and $\alpha_c$ values which are the correlation between the outcome value and whether or not a participant drops out of the study (the magnitude of the data that are MNAR). Specifically, additional increments (starting at .30) were added to the treatment group propensity to respond or $\alpha_t$, and the control group propensity to respond or $\alpha_c$, was increased in increments of .05 for each of the values of the treatment group propensity to respond ($\alpha_t$) without equaling or exceeding the $\alpha_t$ value, starting with the $\alpha_t$ value of .25 and ending with the $\alpha_t$ value of .35.
The simulation model remained the same. The only change in the model was the input values for $\alpha_t$ and $\alpha_c$. The vectors of $\alpha_t$ (propensity to respond in the treatment group) and $\alpha_c$ (propensity to respond in the control group) are:

$$\alpha_t = (25,.25,.25,.25,.30,.30,.30,.30,.30,.35,.35,.35,.35,.35,.35)$$

$$\alpha_c = (05,.10,.15,.20,.05,.10,.15,.20,.25,.05,.10,.15,.20,.25,.30)$$

The vectors $\rho_t$ (response rate of treatment group) and $\rho_c$ (response rate of control group) are (same as original WWC model):

$$\rho_t = (.9, .89, .875, .865, .85, .8, .79, .775, .765, .75, .7, .69, .675, .665, .65, .6, .59, .575, .565, .55)$$

$$\rho_c = (.9, .91, .925, .935, .95, .8, .81, .825, .835, .85, .7, .71, .725, .735, .75, .6, .61, .625, .635, .65)$$

Scenario 5 held the $\alpha_c$ values, which are the correlation between the outcome and whether or not a control group member drops out of the study, to zero and varied the $\alpha_t$ values as in the original WWC attrition model. This scenario modeled a plausible situation in which the control group members are passive and outcome data may be collected on them without them even knowing they are participating in a study. For example, control group participants for an after-school enrichment study may just have their student achievement test data pulled by researchers without the students being aware they are part of a research project. Any missing outcome data on the control group in this situation could plausibly be MCAR or MAR.

The vectors of $\alpha_t$ (propensity to respond in the treatment group) and $\alpha_c$ (propensity to respond in the control group) are:
The vectors $\rho_t$ (response rate of treatment group) and $\rho_c$ (response rate of control group) are (same as original WWC model):

$\rho_t = (\cdot9, .89, .875, .865, .85, .8, .79, .775, .765, .75, .7, .69, .675, .665, .65, .6, .59, .575, .565, .55)$

$\rho_c = (.9, .91, .925, .935, .95, .8, .81, .825, .835, .85, .7, .71, .725, .735, .75, .6, .61, .625, .635, .65)$

Scenario 6 held the $\alpha_c$ values, which are the correlation between the outcome value and whether or not a control group member drops out of the study, to zero and varied the $\alpha_t$ values as in the original WWC attrition model. The scenario also held the response rate of the control group ($\rho_c$) at .90. This scenario modeled a plausible situation similar to that of scenario 5 in which the control group members are passive and outcome data may be collected on them without them even knowing they are participating in a study. For example, attrition in younger students in a school-based study is often relatively low because the only loss of participants is through parents moving out of the district (U.S. Department of Education, 2008, 2011b).

The vectors of $\alpha_t$ (propensity to respond in the treatment group) and $\alpha_c$ (propensity to respond in the control group) are:

$\alpha_t = (.075, 0.1, 0.15, 0.2, 0.3, 0.5, 1)$

$\alpha_c = (0,0,0,0,0,0)$
The vectors $\rho_t$ (response rate of treatment group) and $\rho_c$ (response rate of control group) are:

$\rho_t = (.9, .89, .875, .85, .8, .79, .775, .75, .7, .69, .675, .665, .65, .6, .59, .575, .565, .55)$

$\rho_c = (.9, .9, .9, .9, .9, .9, .9, .9, .9, .9, .9, .9, .9, .9, .9, .9, .9, .9)$

Scenario 7 varied the $\alpha_c$ values as in the original WWC attrition model, and changed the $\alpha_t$ value to equal the $\alpha_c$ values. This scenario modeled a plausible situation in which the relationship between the propensity to respond and the outcome are equal for the treatment and control group, or there is no difference in the propensity to respond between the treatment and control groups. Since WWC reviewers are required to make a judgment about the possible relationship between the propensity to respond and outcome for the treatment and control groups, it is plausible that for certain studies, reviewers could assume the same propensity to respond for both groups.

The vectors of $\alpha_t$ (propensity to respond in the treatment group) and $\alpha_c$ (propensity to respond in the control group) are:

$\alpha_t = (0.075, 0.1, 0.15, 0.2, 0.3, 0.5, 1)$

$\alpha_c = (0.075, 0.1, 0.15, 0.2, 0.3, 0.5, 1)$

The vectors $\rho_t$ (response rate of treatment group) and $\rho_c$ (response rate of control group) are (same as original WWC model):

$\rho_t = (.9, .89, .875, .85, .8, .79, .775, .75, .7, .69, .675, .665, .65, .6, .59, .575, .565, .55)$
Simulation scenario 7 concludes the simulations addressing problem 2.

Problem 3: How does varying the relative magnitude of a specific component of the factor unrelated to attrition in the model (the reliability of the instrument) impact the attrition thresholds? The WWC researchers included in their model a value for \( u \), or a factor unrelated to attrition. This was simulated as a standard normal distribution \( (N[0,1]) \). For most of the studies reviewed by the WWC, an instrument is used to collect outcome data. That instrument, if field tested, should have a reported reliability coefficient. Since the random error term is a component of the factor unrelated to attrition, and since the reliability of the instrument affects the random error, modeling the impact of more or less reliable outcome instruments provides an indication of the impact of instrument reliability on the bias estimates and thresholds set by the WWC. For example, if two studies under review by WWC reviewers are both in the yellow range for gauging acceptable bias based on overall and differential attrition as well as how much the reviewers gauge that the propensity to respond may be related to outcomes (see Figure 1), but the reported reliability of the outcome instrument of one study is .66 and the reported reliability of the outcome instrument for the second study is .91, should reviewers make different decisions about whether or not the studies meet standards for inclusion into the WWC? Scenario 8 models the relative impact across studies of more or less reliable outcome instruments on bias.
Simulation scenario 8- addressing problem 3. Scenario 8 replicated the WWC original inputs for the relationship between the propensity to respond and outcome and the response rates, but modified the model specification for generating the distribution of the $u_t$ (factors unrelated to attrition for the treatment group) and the $u_c$ (factors unrelated to attrition for the control group). The scenario will draw $u_t$ and $u_c$ values from an R-generated normal distribution with a mean of 0 and a standard deviation of 1.2.

Problem 4: Are there discernible patterns to the increase or decrease in bias at different increments of overall and differential attrition and propensity to respond in randomized controlled trials? The final component of this study added value to the overall research question by using data visualization to create a series of heat maps with the matrices produced in the simulation. Heat maps allow for easier consumption of data patterns in matrices by shading cells based on their values (Association for Institutional Research 2013). In Chapter 4, heat maps were used to illustrate the results of each of the 7 simulation scenarios, illustrating in a consumable format how changes in the assumptions related to overall and differential attrition as well as the relationship between the propensity to respond and outcome impact the bias thresholds set by the WWC.
Chapter 4: Results

Research Questions

This dissertation focused on the following research questions:

1. How do the amount and type of attrition, under varying assumptions of how much subjects’ likelihood of dropping out is related to their outcome (MNAR), impact RCTs by contributing to systematic bias, based on the WWC model of bias?

2. How sensitive is the WWC model of bias to changes in attrition rates and/or the relationship between outcome and a participant’s propensity to respond in RCTs?

3. How does varying the random error to simulate variations in the reliability of instruments used across studies impact the attrition thresholds?

4. Are there discernible patterns to the increase or decrease in bias at different increments of overall and differential attrition and propensity to respond in RCTs?

Replicating and Extending What Works Clearinghouse Attrition Bias Model

In order to address the research questions, it was critical to first replicate the Monte Carlo study originally conducted by Mathematica Policy Research on behalf of the U.S. Department of Education and included in the What Works Clearinghouse Procedures and Standards Handbook, Versions 2.0 and 2.1 (U.S. Department of Education, 2008, 2011b). The researchers who conducted the original simulation provided their methodology for creating the distributions as well as their formulas for calculating the bias estimates, but did not provide the R programming code. This dissertation was successful in replicating the Monte Carlo simulation conducted by
Mathematica Policy Research. Table 4 provides the results of the successful replication of the original Monte Carlo study. Comparing the results from Table 4 with the original WWC attrition bias table (Table 2 in this report) it is evident that the results were replicated almost exactly. Recall that the methodology for this dissertation utilized 50,000 replications or draws from the distributions, whereas the original study used only 5,000 replications, perhaps creating differences in the stability of the estimates (Robey & Barcikowski, 1992). Also, this simulation chose an arbitrary seed value for the random number generation that was likely different than the seed value chosen by WWC researchers. Even with the discrepant number of replications between the two studies and a different seed set for random number generation, only three percent or five of the 160 bias estimates generated were consequentially different, or above/below the .05 bias threshold in one study but not the other.

This dissertation was perhaps the first to replicate the original WWC study using the same/similar methodology (at least based on published work). Recall that Vanderbilt methodologists Dong and Lipsey replicated the original WWC attrition bias thresholds, but used a different methodology and a macro program written in SAS 9.2 (Dong & Lipsey, 2010). By using the same methodology as the original researchers, this dissertation both confirmed and then was able to extend the original model of attrition bias.

In order to address Research Question 4 related to visual displays of patterns in the simulation results, Table 4 and the subsequent tables presenting the results of the analyses were produced as a heat table or heat map. A heat map uses colors to represent
cell values and the magnitude of difference between cell values (Association for Institutional Research, 2013). In the heat maps or tables created for this study the green, underlined cells in the matrix indicate that the bias estimate (in effect size units) is at or below the WWC threshold of .05. Bias estimates that fall above .05 are shaded from yellow to dark red, depending on their positive deviation from the .05 threshold (relative deviation specific to each table).

Table 4

Results of Scenario 1 Simulation: Replicating Original Attrition Bias Model

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha_t=0.075$</td>
<td>$\alpha_t=0.10$</td>
<td>$\alpha_t=0.15$</td>
<td>$\alpha_t=0.20$</td>
<td>$\alpha_t=0.30$</td>
<td>$\alpha_t=0.50$</td>
<td>$\alpha_t=1.00$</td>
<td>$\alpha_t=1.00$</td>
<td></td>
</tr>
<tr>
<td>$P_c$ = 0.05</td>
<td>$P_c$ = 0.05</td>
<td>$P_c$ = 0.05</td>
<td>$P_c$ = 0.15</td>
<td>$P_c$ = 0.20</td>
<td>$P_c$ = 0.20</td>
<td>$P_c$ = 1.00</td>
<td>$P_c$ = 1.00</td>
<td></td>
</tr>
<tr>
<td>0.9</td>
<td>0.9</td>
<td>0.1</td>
<td>0.04</td>
<td>0.01</td>
<td>0.02</td>
<td>0.05</td>
<td>-0.01</td>
<td>0.39</td>
</tr>
<tr>
<td>0.89</td>
<td>0.9</td>
<td>0.02</td>
<td>0.03</td>
<td>0.03</td>
<td>0.02</td>
<td>0.03</td>
<td>0.07</td>
<td>0.04</td>
</tr>
<tr>
<td>0.875</td>
<td>0.925</td>
<td>0.03</td>
<td>0.04</td>
<td>0.05</td>
<td>0.04</td>
<td>0.06</td>
<td>0.10</td>
<td>0.09</td>
</tr>
<tr>
<td>0.865</td>
<td>0.935</td>
<td>0.05</td>
<td>0.05</td>
<td>0.06</td>
<td>0.06</td>
<td>0.08</td>
<td>0.11</td>
<td>0.12</td>
</tr>
<tr>
<td>0.85</td>
<td>0.95</td>
<td>0.06</td>
<td>0.06</td>
<td>0.08</td>
<td>0.09</td>
<td>0.10</td>
<td>0.14</td>
<td>0.16</td>
</tr>
<tr>
<td>0.8</td>
<td>0.8</td>
<td>0.01</td>
<td>0.03</td>
<td>0.06</td>
<td>0.03</td>
<td>0.03</td>
<td>0.09</td>
<td>0.02</td>
</tr>
<tr>
<td>0.79</td>
<td>0.81</td>
<td>0.02</td>
<td>0.04</td>
<td>0.06</td>
<td>0.03</td>
<td>0.05</td>
<td>0.12</td>
<td>0.03</td>
</tr>
<tr>
<td>0.775</td>
<td>0.825</td>
<td>0.03</td>
<td>0.06</td>
<td>0.07</td>
<td>0.05</td>
<td>0.08</td>
<td>0.13</td>
<td>0.08</td>
</tr>
<tr>
<td>0.765</td>
<td>0.835</td>
<td>0.04</td>
<td>0.06</td>
<td>0.09</td>
<td>0.07</td>
<td>0.09</td>
<td>0.15</td>
<td>0.10</td>
</tr>
<tr>
<td>0.75</td>
<td>0.85</td>
<td>0.05</td>
<td>0.07</td>
<td>0.10</td>
<td>0.08</td>
<td>0.11</td>
<td>0.18</td>
<td>0.15</td>
</tr>
<tr>
<td>0.7</td>
<td>0.7</td>
<td>0.02</td>
<td>0.05</td>
<td>0.08</td>
<td>0.03</td>
<td>0.04</td>
<td>0.13</td>
<td>0.01</td>
</tr>
<tr>
<td>0.69</td>
<td>0.71</td>
<td>0.03</td>
<td>0.06</td>
<td>0.09</td>
<td>0.05</td>
<td>0.06</td>
<td>0.12</td>
<td>0.02</td>
</tr>
<tr>
<td>0.675</td>
<td>0.725</td>
<td>0.05</td>
<td>0.06</td>
<td>0.11</td>
<td>0.07</td>
<td>0.09</td>
<td>0.18</td>
<td>0.07</td>
</tr>
<tr>
<td>0.665</td>
<td>0.735</td>
<td>0.05</td>
<td>0.08</td>
<td>0.10</td>
<td>0.08</td>
<td>0.10</td>
<td>0.19</td>
<td>0.11</td>
</tr>
<tr>
<td>0.65</td>
<td>0.75</td>
<td>0.06</td>
<td>0.08</td>
<td>0.12</td>
<td>0.09</td>
<td>0.12</td>
<td>0.21</td>
<td>0.15</td>
</tr>
<tr>
<td>0.6</td>
<td>0.6</td>
<td>0.03</td>
<td>0.06</td>
<td>0.10</td>
<td>0.04</td>
<td>0.06</td>
<td>0.17</td>
<td>0.00</td>
</tr>
<tr>
<td>0.59</td>
<td>0.61</td>
<td>0.05</td>
<td>0.08</td>
<td>0.12</td>
<td>0.05</td>
<td>0.07</td>
<td>0.19</td>
<td>0.03</td>
</tr>
<tr>
<td>0.575</td>
<td>0.625</td>
<td>0.06</td>
<td>0.08</td>
<td>0.13</td>
<td>0.07</td>
<td>0.10</td>
<td>0.21</td>
<td>0.07</td>
</tr>
<tr>
<td>0.565</td>
<td>0.635</td>
<td>0.06</td>
<td>0.08</td>
<td>0.14</td>
<td>0.08</td>
<td>0.12</td>
<td>0.22</td>
<td>0.10</td>
</tr>
<tr>
<td>0.55</td>
<td>0.65</td>
<td>0.07</td>
<td>0.10</td>
<td>0.16</td>
<td>0.10</td>
<td>0.15</td>
<td>0.26</td>
<td>0.16</td>
</tr>
</tbody>
</table>

Green cells meet the .05 bias threshold
Table 5 presents the results of Scenario 2, further validating the WWC model by setting the relationship between the propensity to respond and the outcome (the $\alpha$ term) to 0 and (in Row 1) setting the attrition to 0 for both the treatment group and the control group. Rows 2 through 21 allowed the overall and differential attrition to vary in the same increments as in the original WWC study but still kept the relationship between the propensity to respond and outcome at zero.

As expected, the bias estimates were all at zero or near zero as the WWC formula is driven by the difference in mean bias between the treatment group and the control group. Recall that bias estimates in the WWC model are the differences between the mean outcomes ($y$) of the treatment group respondents and the control group respondents. It is driven by the proportion of the variation in $y$ explained by the response rates and how much the propensity to respond is related to the outcome (U.S. Department of Education, 2008, 2011b). The basic formula is:

$$y = \alpha * z + (1-\alpha)u$$

According to the formula, if the $\alpha$ or relationship between the propensity to respond and outcome is set to zero, then only random differences (the $u$, or factors unrelated to attrition) in $y_t$ and $y_c$ exist. That computational formula mimics the assertion in the literature on attrition and bias presented in Chapter 1 of this dissertation—that, although all attrition is unwanted because it may impact statistical power, outcome data that are
MCAR or MAR do not necessarily introduce bias or pose a substantial threat to validity (Bell et al., 2013; Fitzgerald et al., 1997).

Table 5

*Results of Scenario 2 Simulation: Modeling the Counterfactual*

<table>
<thead>
<tr>
<th>$P_t$</th>
<th>$P_c$</th>
<th>$\alpha_c=0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0.00</td>
</tr>
<tr>
<td>0.9</td>
<td>0.9</td>
<td>0.00</td>
</tr>
<tr>
<td>0.89</td>
<td>0.91</td>
<td>0.01</td>
</tr>
<tr>
<td>0.875</td>
<td>0.925</td>
<td>0.00</td>
</tr>
<tr>
<td>0.865</td>
<td>0.935</td>
<td>0.00</td>
</tr>
<tr>
<td>0.85</td>
<td>0.95</td>
<td>0.00</td>
</tr>
<tr>
<td>0.8</td>
<td>0.8</td>
<td>0.00</td>
</tr>
<tr>
<td>0.79</td>
<td>0.81</td>
<td>-0.01</td>
</tr>
<tr>
<td>0.775</td>
<td>0.825</td>
<td>0.00</td>
</tr>
<tr>
<td>0.765</td>
<td>0.835</td>
<td>0.00</td>
</tr>
<tr>
<td>0.75</td>
<td>0.85</td>
<td>-0.01</td>
</tr>
<tr>
<td>0.7</td>
<td>0.7</td>
<td>-0.01</td>
</tr>
<tr>
<td>0.69</td>
<td>0.71</td>
<td>0.00</td>
</tr>
<tr>
<td>0.675</td>
<td>0.725</td>
<td>0.02</td>
</tr>
<tr>
<td>0.665</td>
<td>0.735</td>
<td>-0.01</td>
</tr>
<tr>
<td>0.65</td>
<td>0.75</td>
<td>0.00</td>
</tr>
<tr>
<td>0.6</td>
<td>0.6</td>
<td>-0.01</td>
</tr>
<tr>
<td>0.59</td>
<td>0.61</td>
<td>-0.01</td>
</tr>
<tr>
<td>0.575</td>
<td>0.625</td>
<td>0.00</td>
</tr>
<tr>
<td>0.565</td>
<td>0.635</td>
<td>0.00</td>
</tr>
<tr>
<td>0.55</td>
<td>0.65</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Table 6 and Table 7 begin the sensitivity analysis of the WWC attrition bias model by more systematically varying the $\alpha_t$ and $\alpha_c$ values which are the correlation between the outcome value and whether or not a participant drops out of the study (the magnitude of the data that are MNAR). Specifically, more increments were added to the
treatment group propensity to respond, or \( \alpha_t \). The control group propensity to respond, or \( \alpha_c \), was then increased in increments of .05 for each of the values of the treatment group propensity to respond (\( \alpha_t \)) without equaling or exceeding the \( \alpha_t \) value, starting with the \( \alpha_t \) value of .075 and ending with the \( \alpha_t \) value of .20 for Table 6, and starting with the \( \alpha_t \) value of .25 for Table 7.

As the Table 6 heat chart indicates visually, the difference in the magnitude of the relationship between the propensity to respond and outcome (\( \alpha_t \) and \( \alpha_c \)) drive the bias estimates up, largely regardless of the overall magnitude of the relationship between the propensity to respond and outcome, and largely regardless of the amount of overall and differential attrition. For example, where the \( \alpha_t \) and \( \alpha_c \) were set at equal or nearly equal (columns 2, 6, and 8) the bias estimates remain under the acceptable threshold regardless of the overall or differential attrition rates. Where the \( \alpha_t \) and \( \alpha_c \) were set at most divergent (columns 9 and 11) the bias estimates are largely over the threshold regardless of overall or differential attrition. While this is intuitive by dissecting the bias formula, the original WWC bias study does not illustrate that clearly in the table of bias estimates (Table 2 in this report).

Table 7 illustrates the higher overall magnitude of relationship of the propensity to respond to outcome (.25 through .35). As the heat chart indicates, while fewer cells have values that fall under the .05 bias threshold than in Table 6, the columns with the more divergent \( \alpha_t \) and \( \alpha_c \) values (columns 1, 5, and 10) show the darkest red, or highest bias estimates and those with the least divergent \( \alpha_t \) and \( \alpha_c \) values show more cells with values under the .05 bias threshold.
Table 6

Results of Scenario 3 Simulation: Systematically Extending Alphas for Sensitivity Analysis

<table>
<thead>
<tr>
<th>$P_t$</th>
<th>$P_c$</th>
<th>$\alpha_t=0.075$</th>
<th>$\alpha_t=0.075$</th>
<th>$\alpha_t=0.085$</th>
<th>$\alpha_t=0.085$</th>
<th>$\alpha_t=0.095$</th>
<th>$\alpha_t=0.095$</th>
<th>$\alpha_t=0.10$</th>
<th>$\alpha_t=0.10$</th>
<th>$\alpha_t=0.15$</th>
<th>$\alpha_t=0.15$</th>
<th>$\alpha_t=0.20$</th>
<th>$\alpha_t=0.20$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.9</td>
<td>0.9</td>
<td>0.01</td>
<td>-0.01</td>
<td>0.02</td>
<td>-0.01</td>
<td>0.02</td>
<td>0.00</td>
<td>0.01</td>
<td>-0.01</td>
<td>0.03</td>
<td>0.02</td>
<td>0.04</td>
<td>0.02</td>
</tr>
<tr>
<td>0.89</td>
<td>0.91</td>
<td>0.02</td>
<td>0.00</td>
<td>0.02</td>
<td>0.01</td>
<td>0.03</td>
<td>0.01</td>
<td>0.03</td>
<td>0.02</td>
<td>0.04</td>
<td>0.02</td>
<td>0.05</td>
<td>0.02</td>
</tr>
<tr>
<td>0.875</td>
<td>0.925</td>
<td>0.03</td>
<td>0.02</td>
<td>0.04</td>
<td>0.02</td>
<td>0.04</td>
<td>0.03</td>
<td>0.03</td>
<td>0.03</td>
<td>0.06</td>
<td>0.05</td>
<td>0.07</td>
<td>0.05</td>
</tr>
<tr>
<td>0.865</td>
<td>0.935</td>
<td>0.04</td>
<td>0.02</td>
<td>0.05</td>
<td>0.03</td>
<td>0.05</td>
<td>0.03</td>
<td>0.05</td>
<td>0.03</td>
<td>0.06</td>
<td>0.06</td>
<td>0.08</td>
<td>0.07</td>
</tr>
<tr>
<td>0.85</td>
<td>0.95</td>
<td>0.06</td>
<td>0.05</td>
<td>0.06</td>
<td>0.06</td>
<td>0.06</td>
<td>0.05</td>
<td>0.06</td>
<td>0.05</td>
<td>0.08</td>
<td>0.07</td>
<td>0.10</td>
<td>0.09</td>
</tr>
<tr>
<td>0.8</td>
<td>0.8</td>
<td>0.02</td>
<td>-0.01</td>
<td>0.01</td>
<td>-0.01</td>
<td>0.04</td>
<td>0.00</td>
<td>0.04</td>
<td>-0.01</td>
<td>0.06</td>
<td>0.02</td>
<td>0.07</td>
<td>0.05</td>
</tr>
<tr>
<td>0.79</td>
<td>0.81</td>
<td>0.03</td>
<td>-0.01</td>
<td>0.03</td>
<td>0.01</td>
<td>0.03</td>
<td>0.01</td>
<td>0.04</td>
<td>0.01</td>
<td>0.07</td>
<td>0.04</td>
<td>0.09</td>
<td>0.06</td>
</tr>
<tr>
<td>0.775</td>
<td>0.825</td>
<td>0.04</td>
<td>0.01</td>
<td>0.04</td>
<td>0.01</td>
<td>0.05</td>
<td>0.02</td>
<td>0.05</td>
<td>0.03</td>
<td>0.08</td>
<td>0.06</td>
<td>0.11</td>
<td>0.07</td>
</tr>
<tr>
<td>0.765</td>
<td>0.835</td>
<td>0.05</td>
<td>0.02</td>
<td>0.05</td>
<td>0.03</td>
<td>0.06</td>
<td>0.03</td>
<td>0.06</td>
<td>0.03</td>
<td>0.09</td>
<td>0.06</td>
<td>0.12</td>
<td>0.08</td>
</tr>
<tr>
<td>0.75</td>
<td>0.85</td>
<td>0.06</td>
<td>0.03</td>
<td>0.07</td>
<td>0.04</td>
<td>0.08</td>
<td>0.05</td>
<td>0.08</td>
<td>0.05</td>
<td>0.10</td>
<td>0.07</td>
<td>0.13</td>
<td>0.10</td>
</tr>
<tr>
<td>0.7</td>
<td>0.7</td>
<td>0.03</td>
<td>-0.03</td>
<td>0.04</td>
<td>-0.02</td>
<td>0.04</td>
<td>-0.01</td>
<td>0.04</td>
<td>-0.01</td>
<td>0.09</td>
<td>0.04</td>
<td>0.11</td>
<td>0.07</td>
</tr>
<tr>
<td>0.69</td>
<td>0.71</td>
<td>0.02</td>
<td>-0.01</td>
<td>0.05</td>
<td>-0.01</td>
<td>0.05</td>
<td>0.00</td>
<td>0.05</td>
<td>0.00</td>
<td>0.09</td>
<td>0.05</td>
<td>0.13</td>
<td>0.07</td>
</tr>
<tr>
<td>0.675</td>
<td>0.725</td>
<td>0.04</td>
<td>0.00</td>
<td>0.05</td>
<td>0.01</td>
<td>0.06</td>
<td>0.01</td>
<td>0.07</td>
<td>0.01</td>
<td>0.11</td>
<td>0.06</td>
<td>0.14</td>
<td>0.10</td>
</tr>
<tr>
<td>0.665</td>
<td>0.735</td>
<td>0.05</td>
<td>0.00</td>
<td>0.05</td>
<td>0.02</td>
<td>0.07</td>
<td>0.03</td>
<td>0.07</td>
<td>0.03</td>
<td>0.11</td>
<td>0.08</td>
<td>0.15</td>
<td>0.10</td>
</tr>
<tr>
<td>0.65</td>
<td>0.75</td>
<td>0.06</td>
<td>0.02</td>
<td>0.08</td>
<td>0.04</td>
<td>0.08</td>
<td>0.05</td>
<td>0.09</td>
<td>0.05</td>
<td>0.13</td>
<td>0.09</td>
<td>0.16</td>
<td>0.12</td>
</tr>
<tr>
<td>0.6</td>
<td>0.6</td>
<td>0.04</td>
<td>-0.03</td>
<td>0.05</td>
<td>-0.01</td>
<td>0.06</td>
<td>-0.01</td>
<td>0.05</td>
<td>0.00</td>
<td>0.11</td>
<td>0.06</td>
<td>0.15</td>
<td>0.07</td>
</tr>
<tr>
<td>0.59</td>
<td>0.61</td>
<td>0.03</td>
<td>-0.02</td>
<td>0.06</td>
<td>0.00</td>
<td>0.06</td>
<td>0.00</td>
<td>0.06</td>
<td>0.01</td>
<td>0.13</td>
<td>0.06</td>
<td>0.16</td>
<td>0.09</td>
</tr>
<tr>
<td>0.575</td>
<td>0.625</td>
<td>0.05</td>
<td>-0.01</td>
<td>0.07</td>
<td>0.00</td>
<td>0.06</td>
<td>0.02</td>
<td>0.07</td>
<td>0.02</td>
<td>0.13</td>
<td>0.08</td>
<td>0.17</td>
<td>0.11</td>
</tr>
<tr>
<td>0.565</td>
<td>0.635</td>
<td>0.07</td>
<td>-0.01</td>
<td>0.09</td>
<td>0.02</td>
<td>0.08</td>
<td>0.03</td>
<td>0.10</td>
<td>0.03</td>
<td>0.15</td>
<td>0.09</td>
<td>0.17</td>
<td>0.14</td>
</tr>
<tr>
<td>0.55</td>
<td>0.65</td>
<td>0.07</td>
<td>0.02</td>
<td>0.08</td>
<td>0.03</td>
<td>0.09</td>
<td>0.04</td>
<td>0.11</td>
<td>0.05</td>
<td>0.15</td>
<td>0.10</td>
<td>0.19</td>
<td>0.15</td>
</tr>
</tbody>
</table>

Green cells meet the .05 bias threshold.
## Table 7

**Results of Scenario 4 Simulation: Systematically Extending Alphas for Sensitivity Analysis, Part 2**

<table>
<thead>
<tr>
<th>P_i</th>
<th>P_c</th>
<th>α_t=0.25</th>
<th>α_t=0.25</th>
<th>α_t=0.30</th>
<th>α_t=0.30</th>
<th>α_t=0.30</th>
<th>α_t=0.30</th>
<th>α_t=0.35</th>
<th>α_t=0.35</th>
<th>α_t=0.35</th>
<th>α_t=0.35</th>
<th>α_t=0.35</th>
<th>α_t=0.35</th>
<th>α_t=0.35</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.9</td>
<td>0.9</td>
<td>0.05</td>
<td>0.04</td>
<td>0.02</td>
<td>0.01</td>
<td>0.06</td>
<td>0.04</td>
<td>0.02</td>
<td>0.01</td>
<td>0.07</td>
<td>0.05</td>
<td>0.04</td>
<td>0.03</td>
<td>0.02</td>
</tr>
<tr>
<td>0.89</td>
<td>0.91</td>
<td>0.07</td>
<td>0.05</td>
<td>0.04</td>
<td>0.03</td>
<td>0.07</td>
<td>0.05</td>
<td>0.04</td>
<td>0.03</td>
<td>0.08</td>
<td>0.07</td>
<td>0.06</td>
<td>0.04</td>
<td>0.03</td>
</tr>
<tr>
<td>0.875</td>
<td>0.925</td>
<td>0.09</td>
<td>0.07</td>
<td>0.06</td>
<td>0.04</td>
<td>0.10</td>
<td>0.08</td>
<td>0.06</td>
<td>0.05</td>
<td>0.11</td>
<td>0.09</td>
<td>0.07</td>
<td>0.08</td>
<td>0.06</td>
</tr>
<tr>
<td>0.865</td>
<td>0.935</td>
<td>0.09</td>
<td>0.08</td>
<td>0.07</td>
<td>0.06</td>
<td>0.11</td>
<td>0.09</td>
<td>0.10</td>
<td>0.08</td>
<td>0.07</td>
<td>0.12</td>
<td>0.11</td>
<td>0.09</td>
<td>0.08</td>
</tr>
<tr>
<td>0.85</td>
<td>0.95</td>
<td>0.11</td>
<td>0.10</td>
<td>0.10</td>
<td>0.08</td>
<td>0.13</td>
<td>0.11</td>
<td>0.11</td>
<td>0.10</td>
<td>0.13</td>
<td>0.12</td>
<td>0.12</td>
<td>0.11</td>
<td>0.10</td>
</tr>
<tr>
<td>0.8</td>
<td>0.8</td>
<td>0.10</td>
<td>0.07</td>
<td>0.04</td>
<td>0.02</td>
<td>0.12</td>
<td>0.08</td>
<td>0.06</td>
<td>0.03</td>
<td>0.13</td>
<td>0.09</td>
<td>0.07</td>
<td>0.05</td>
<td>0.03</td>
</tr>
<tr>
<td>0.79</td>
<td>0.81</td>
<td>0.11</td>
<td>0.08</td>
<td>0.05</td>
<td>0.03</td>
<td>0.13</td>
<td>0.10</td>
<td>0.08</td>
<td>0.06</td>
<td>0.15</td>
<td>0.11</td>
<td>0.08</td>
<td>0.06</td>
<td>0.05</td>
</tr>
<tr>
<td>0.775</td>
<td>0.825</td>
<td>0.11</td>
<td>0.10</td>
<td>0.07</td>
<td>0.05</td>
<td>0.14</td>
<td>0.11</td>
<td>0.09</td>
<td>0.07</td>
<td>0.16</td>
<td>0.13</td>
<td>0.10</td>
<td>0.09</td>
<td>0.08</td>
</tr>
<tr>
<td>0.765</td>
<td>0.835</td>
<td>0.13</td>
<td>0.11</td>
<td>0.08</td>
<td>0.07</td>
<td>0.16</td>
<td>0.13</td>
<td>0.11</td>
<td>0.09</td>
<td>0.17</td>
<td>0.14</td>
<td>0.12</td>
<td>0.11</td>
<td>0.09</td>
</tr>
<tr>
<td>0.75</td>
<td>0.85</td>
<td>0.15</td>
<td>0.12</td>
<td>0.11</td>
<td>0.09</td>
<td>0.18</td>
<td>0.15</td>
<td>0.13</td>
<td>0.11</td>
<td>0.19</td>
<td>0.17</td>
<td>0.14</td>
<td>0.14</td>
<td>0.11</td>
</tr>
<tr>
<td>0.7</td>
<td>0.7</td>
<td>0.14</td>
<td>0.09</td>
<td>0.06</td>
<td>0.03</td>
<td>0.16</td>
<td>0.12</td>
<td>0.08</td>
<td>0.05</td>
<td>0.18</td>
<td>0.14</td>
<td>0.10</td>
<td>0.07</td>
<td>0.05</td>
</tr>
<tr>
<td>0.69</td>
<td>0.71</td>
<td>0.14</td>
<td>0.10</td>
<td>0.07</td>
<td>0.04</td>
<td>0.17</td>
<td>0.13</td>
<td>0.10</td>
<td>0.06</td>
<td>0.19</td>
<td>0.14</td>
<td>0.12</td>
<td>0.09</td>
<td>0.06</td>
</tr>
<tr>
<td>0.675</td>
<td>0.725</td>
<td>0.16</td>
<td>0.12</td>
<td>0.09</td>
<td>0.07</td>
<td>0.18</td>
<td>0.15</td>
<td>0.11</td>
<td>0.09</td>
<td>0.21</td>
<td>0.17</td>
<td>0.14</td>
<td>0.11</td>
<td>0.08</td>
</tr>
<tr>
<td>0.665</td>
<td>0.735</td>
<td>0.18</td>
<td>0.13</td>
<td>0.10</td>
<td>0.08</td>
<td>0.20</td>
<td>0.16</td>
<td>0.12</td>
<td>0.10</td>
<td>0.23</td>
<td>0.17</td>
<td>0.15</td>
<td>0.12</td>
<td>0.10</td>
</tr>
<tr>
<td>0.65</td>
<td>0.75</td>
<td>0.19</td>
<td>0.14</td>
<td>0.12</td>
<td>0.10</td>
<td>0.22</td>
<td>0.18</td>
<td>0.15</td>
<td>0.13</td>
<td>0.24</td>
<td>0.21</td>
<td>0.17</td>
<td>0.14</td>
<td>0.13</td>
</tr>
<tr>
<td>0.6</td>
<td>0.6</td>
<td>0.18</td>
<td>0.11</td>
<td>0.07</td>
<td>0.03</td>
<td>0.21</td>
<td>0.15</td>
<td>0.10</td>
<td>0.06</td>
<td>0.24</td>
<td>0.17</td>
<td>0.14</td>
<td>0.10</td>
<td>0.06</td>
</tr>
<tr>
<td>0.59</td>
<td>0.61</td>
<td>0.19</td>
<td>0.12</td>
<td>0.08</td>
<td>0.05</td>
<td>0.21</td>
<td>0.15</td>
<td>0.13</td>
<td>0.08</td>
<td>0.26</td>
<td>0.19</td>
<td>0.15</td>
<td>0.11</td>
<td>0.08</td>
</tr>
<tr>
<td>0.575</td>
<td>0.625</td>
<td>0.21</td>
<td>0.15</td>
<td>0.11</td>
<td>0.07</td>
<td>0.23</td>
<td>0.18</td>
<td>0.14</td>
<td>0.10</td>
<td>0.26</td>
<td>0.21</td>
<td>0.16</td>
<td>0.13</td>
<td>0.10</td>
</tr>
<tr>
<td>0.565</td>
<td>0.635</td>
<td>0.21</td>
<td>0.15</td>
<td>0.12</td>
<td>0.09</td>
<td>0.24</td>
<td>0.20</td>
<td>0.15</td>
<td>0.11</td>
<td>0.28</td>
<td>0.22</td>
<td>0.18</td>
<td>0.15</td>
<td>0.12</td>
</tr>
<tr>
<td>0.55</td>
<td>0.65</td>
<td>0.23</td>
<td>0.17</td>
<td>0.14</td>
<td>0.11</td>
<td>0.26</td>
<td>0.22</td>
<td>0.17</td>
<td>0.14</td>
<td>0.30</td>
<td>0.25</td>
<td>0.20</td>
<td>0.17</td>
<td>0.15</td>
</tr>
</tbody>
</table>
Table 8 illustrates the results of Scenario 5, which held the $\alpha_c$ values or the relationship between the outcome and whether or not a control group member drops out of the study, to zero and varied the $\alpha_t$ values as in the original WWC attrition model. This scenario modeled a plausible situation in which the control group members are passive and outcome data may be collected on them without them even knowing they are participating in a study. For example, control group participants for an after-school enrichment study may just have their student achievement test data pulled by researchers without the students being aware they are part of a research project. Any missing outcome data on the control group in this situation could plausibly be MCAR or MAR.

As the heat table indicates at first glance, the scenario allowed only one cell to indicate results below the .05 threshold. This finding is driven by the sensitivity of the model to the divergence of the $\alpha_t$ and $\alpha_c$ values rather than the overall magnitude of these values. As Table 8, column 1 indicates, this scenario inflated the bias to intolerable levels even with a relatively low relationship between the propensity to respond and outcome for the treatment group of .075 (the WWC Procedures and Standards Handbook refers to this relationship as “the most optimistic assumption” about the relationship between the propensity to respond and outcome.”) (U.S. Department of Education, 2011, p. 34)
Table 8

Results of Scenario 5 Simulation: Setting Control Group Relationship of The propensity to respond to Outcome at Zero

| P_t | P_c | α_t=0.075 | α_t=0.10 | α_t=0.15 | α_t=0.20 | α_t=0.30 | α_t=0.50 | α_t=1.00 |
|-----|-----|-------|-------|-------|-------|-------|-------|-------|-------|
| 0.9 | 0.9 | 0.05  | 0.06  | 0.09  | 0.11  | 0.14  | 0.19  |       |       |
| 0.89| 0.91| 0.06  | 0.07  | 0.09  | 0.10  | 0.13  | 0.17  | 0.22  | 0.23  |
| 0.875| 0.925| 0.06 | 0.07 | 0.09 | 0.10 | 0.13 | 0.17 | 0.23 |       |
| 0.865| 0.935| 0.08 | 0.08 | 0.09 | 0.11 | 0.14 | 0.17 | 0.26 |       |
| 0.85| 0.95| 0.08 | 0.08 | 0.10 | 0.13 | 0.15 | 0.19 | 0.28 |       |
| 0.8 | 0.8| 0.09 | 0.11 | 0.14 | 0.17 | 0.19 | 0.24 | 0.36 |       |
| 0.79| 0.81| 0.10 | 0.11 | 0.14 | 0.17 | 0.20 | 0.27 | 0.37 |       |
| 0.775| 0.825| 0.10 | 0.12 | 0.14 | 0.17 | 0.22 | 0.27 | 0.39 |       |
| 0.765| 0.835| 0.16 | 0.18 | 0.23 | 0.28 | 0.40 |       |       |       |
| 0.75| 0.85| 0.16 | 0.17 | 0.19 | 0.23 | 0.30 | 0.42 |       |       |
| 0.7 | 0.7| 0.13 | 0.16 | 0.20 | 0.26 | 0.36 | 0.50 |       |       |
| 0.69| 0.71| 0.14 | 0.17 | 0.24 | 0.27 | 0.36 | 0.51 |       |       |
| 0.675| 0.725| 0.21 | 0.25 | 0.30 | 0.39 | 0.53 |       |       |       |
| 0.665| 0.735| 0.25 | 0.30 | 0.39 | 0.56 |       |       |       |       |
| 0.65| 0.75| 0.25 | 0.31 | 0.41 | 0.57 |       |       |       |       |
| 0.6 | 0.6| 0.17 | 0.20 | 0.25 | 0.29 | 0.35 | 0.45 | 0.64 |       |
| 0.59| 0.61| 0.19 | 0.22 | 0.26 | 0.30 | 0.35 | 0.46 | 0.66 |       |
| 0.575| 0.625| 0.27 | 0.31 | 0.37 | 0.48 | 0.69 |       |       |       |
| 0.565| 0.635| 0.27 | 0.31 | 0.39 | 0.48 | 0.69 |       |       |       |
| 0.55| 0.65| 0.28 | 0.33 | 0.40 | 0.51 | 0.73 |       |       |       |

Green cells meet the .05 bias threshold

Scenario 6 (Table 9) held the α_c values, which are the correlation between the outcome value and whether or not a control group member drops out of the study, to zero and varied the α_t values as in the original WWC attrition model. The scenario also held the response rate of the control group (ρ_c) at .90. This scenario modeled a plausible situation similar to that of Scenario 5 in which the control group members are passive
and outcome data may be collected on them without them even knowing they are participating in a study. For example, attrition in younger students in a school-based study is often relatively low because the only loss of participants is through parents moving out of the district (U.S. Department of Education, 2008, 2011b). As indicated by the heat table in Table 9, results of this scenario were similar to those of Scenario 5. The change in the overall and differential response rate had no impact on the magnitude of bias, which is driven primarily by the differential relationship of the propensity to respond to outcome between the treatment and control group.
### Table 9

**Results of Scenario 6 Simulation: Holding Control Group Relationship of The propensity to respond to Outcome at Zero and Holding Control Group Attrition Rate at 10 Percent**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>(P_t)</td>
<td>(P_c)</td>
<td>(α_c=0)</td>
<td>(α_c=0)</td>
<td>(α_c=0)</td>
<td>(α_c=0)</td>
<td>(α_c=0)</td>
<td>(α_c=0)</td>
</tr>
<tr>
<td>0.9</td>
<td>0.9</td>
<td>0.05</td>
<td>0.06</td>
<td>0.09</td>
<td>0.09</td>
<td>0.11</td>
<td>0.14</td>
</tr>
<tr>
<td>0.89</td>
<td>0.9</td>
<td>0.05</td>
<td>0.07</td>
<td>0.08</td>
<td>0.09</td>
<td>0.11</td>
<td>0.14</td>
</tr>
<tr>
<td>0.875</td>
<td>0.9</td>
<td>0.07</td>
<td>0.08</td>
<td>0.09</td>
<td>0.11</td>
<td>0.13</td>
<td>0.16</td>
</tr>
<tr>
<td>0.865</td>
<td>0.9</td>
<td>0.07</td>
<td>0.08</td>
<td>0.10</td>
<td>0.12</td>
<td>0.14</td>
<td>0.17</td>
</tr>
<tr>
<td>0.85</td>
<td>0.9</td>
<td>0.07</td>
<td>0.08</td>
<td>0.11</td>
<td>0.12</td>
<td>0.15</td>
<td>0.19</td>
</tr>
<tr>
<td>0.8</td>
<td>0.9</td>
<td>0.10</td>
<td>0.11</td>
<td>0.13</td>
<td>0.16</td>
<td>0.20</td>
<td>0.24</td>
</tr>
<tr>
<td>0.79</td>
<td>0.9</td>
<td>0.10</td>
<td>0.12</td>
<td>0.14</td>
<td>0.16</td>
<td>0.21</td>
<td>0.27</td>
</tr>
<tr>
<td>0.775</td>
<td>0.9</td>
<td>0.10</td>
<td>0.12</td>
<td>0.15</td>
<td>0.17</td>
<td>0.22</td>
<td>0.28</td>
</tr>
<tr>
<td>0.765</td>
<td>0.9</td>
<td>0.11</td>
<td>0.13</td>
<td>0.14</td>
<td>0.18</td>
<td>0.23</td>
<td>0.28</td>
</tr>
<tr>
<td>0.75</td>
<td>0.9</td>
<td>0.12</td>
<td>0.13</td>
<td>0.16</td>
<td>0.19</td>
<td>0.23</td>
<td>0.31</td>
</tr>
<tr>
<td>0.7</td>
<td>0.9</td>
<td>0.14</td>
<td>0.17</td>
<td>0.19</td>
<td>0.22</td>
<td>0.27</td>
<td>0.35</td>
</tr>
<tr>
<td>0.69</td>
<td>0.9</td>
<td>0.14</td>
<td>0.16</td>
<td>0.19</td>
<td>0.22</td>
<td>0.29</td>
<td>0.37</td>
</tr>
<tr>
<td>0.675</td>
<td>0.9</td>
<td>0.13</td>
<td>0.18</td>
<td>0.21</td>
<td>0.22</td>
<td>0.29</td>
<td>0.39</td>
</tr>
<tr>
<td>0.665</td>
<td>0.9</td>
<td>0.16</td>
<td>0.16</td>
<td>0.21</td>
<td>0.25</td>
<td>0.30</td>
<td>0.39</td>
</tr>
<tr>
<td>0.65</td>
<td>0.9</td>
<td>0.16</td>
<td>0.19</td>
<td>0.22</td>
<td>0.26</td>
<td>0.31</td>
<td>0.40</td>
</tr>
<tr>
<td>0.6</td>
<td>0.9</td>
<td>0.19</td>
<td>0.20</td>
<td>0.25</td>
<td>0.28</td>
<td>0.35</td>
<td>0.46</td>
</tr>
<tr>
<td>0.59</td>
<td>0.9</td>
<td>0.17</td>
<td>0.20</td>
<td>0.25</td>
<td>0.29</td>
<td>0.36</td>
<td>0.47</td>
</tr>
<tr>
<td>0.575</td>
<td>0.9</td>
<td>0.19</td>
<td>0.21</td>
<td>0.26</td>
<td>0.30</td>
<td>0.37</td>
<td>0.48</td>
</tr>
<tr>
<td>0.565</td>
<td>0.9</td>
<td>0.19</td>
<td>0.22</td>
<td>0.28</td>
<td>0.32</td>
<td>0.38</td>
<td>0.50</td>
</tr>
<tr>
<td>0.55</td>
<td>0.9</td>
<td>0.20</td>
<td>0.22</td>
<td>0.28</td>
<td>0.33</td>
<td>0.39</td>
<td>0.51</td>
</tr>
</tbody>
</table>

Green cells meet the .05 bias threshold

Scenario 7 (Table 10) varied the \(\alpha_c\) values as in the original WWC attrition model, and changed the \(\alpha_t\) value to equal the \(\alpha_c\) values. This scenario modeled a plausible situation in which the relationship between the propensity to respond and the outcome are equal for the treatment and control group, or there is no difference in the propensity to respond between the treatment and control groups. Since WWC reviewers are required to...
make a judgment about the possible relationship between the propensity to respond and outcome for the treatment and control groups, it is plausible that for certain studies, reviewers could assume the same propensity to respond for both groups. For example, if a study was conducted in a school and on younger children and was not intrusive, one could reasonably assume that the relationship between the propensity to respond and outcome was equal for both the treatment and control groups.

As the heat chart pattern indicates, the lack of any differential relationship between the propensity to respond and outcome drove the model results (as indicated in earlier scenarios where the $\alpha_t$ and $\alpha_c$ values were close to equal). The heat chart also indicates a clear pattern of cells with bias values over the .05 threshold. Those cells are where the differential attrition rates ($P_t$ and $P_c$) were highest.
Table 10

Results of Simulation Scenario 7: Setting the Participant Group and Control Group Relationship of The propensity to respond to Outcome as Equal

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\alpha_t=0.075$</td>
<td>$\alpha_t=0.10$</td>
<td>$\alpha_t=0.15$</td>
<td>$\alpha_t=0.20$</td>
<td>$\alpha_t=0.30$</td>
<td>$\alpha_t=0.50$</td>
<td>$\alpha_t=1.00$</td>
</tr>
<tr>
<td>$P_t$</td>
<td>0.9</td>
<td>0.9</td>
<td>0.00</td>
<td>0.00</td>
<td>0.01</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>0.89</td>
<td>0.9</td>
<td>0.00</td>
<td>0.01</td>
<td>0.02</td>
<td>0.01</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>0.875</td>
<td>0.925</td>
<td>0.03</td>
<td>0.03</td>
<td>0.03</td>
<td>0.04</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td>0.865</td>
<td>0.935</td>
<td>0.03</td>
<td>0.04</td>
<td>0.04</td>
<td>0.06</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>0.85</td>
<td>0.95</td>
<td>0.04</td>
<td>0.05</td>
<td>0.07</td>
<td>0.07</td>
<td>0.10</td>
</tr>
<tr>
<td></td>
<td>0.8</td>
<td>0.8</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>0.79</td>
<td>0.81</td>
<td>0.01</td>
<td>0.02</td>
<td>0.01</td>
<td>0.01</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td>0.775</td>
<td>0.825</td>
<td>0.02</td>
<td>0.02</td>
<td>0.03</td>
<td>0.03</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td>0.765</td>
<td>0.835</td>
<td>0.03</td>
<td>0.04</td>
<td>0.03</td>
<td>0.05</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>0.75</td>
<td>0.85</td>
<td>0.04</td>
<td>0.04</td>
<td>0.06</td>
<td>0.07</td>
<td>0.08</td>
</tr>
<tr>
<td></td>
<td>0.7</td>
<td>0.7</td>
<td>0.01</td>
<td>0.01</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>0.69</td>
<td>0.71</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>0.675</td>
<td>0.725</td>
<td>0.01</td>
<td>0.03</td>
<td>0.03</td>
<td>0.02</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td>0.665</td>
<td>0.735</td>
<td>0.04</td>
<td>0.02</td>
<td>0.04</td>
<td>0.05</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>0.65</td>
<td>0.75</td>
<td>0.05</td>
<td>0.06</td>
<td>0.06</td>
<td>0.07</td>
<td>0.08</td>
</tr>
<tr>
<td></td>
<td>0.6</td>
<td>0.6</td>
<td>0.01</td>
<td>-0.01</td>
<td>0.00</td>
<td>0.00</td>
<td>-0.01</td>
</tr>
<tr>
<td></td>
<td>0.59</td>
<td>0.61</td>
<td>0.00</td>
<td>0.00</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>0.575</td>
<td>0.625</td>
<td>0.02</td>
<td>0.02</td>
<td>0.03</td>
<td>0.03</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td>0.565</td>
<td>0.635</td>
<td>0.03</td>
<td>0.03</td>
<td>0.05</td>
<td>0.05</td>
<td>0.06</td>
</tr>
<tr>
<td>0.55</td>
<td>0.65</td>
<td>0.05</td>
<td>0.04</td>
<td>0.06</td>
<td>0.07</td>
<td>0.08</td>
<td>0.10</td>
</tr>
</tbody>
</table>

Green cells meet the .05 bias threshold

Scenario 8 (Table 11) replicated the WWC’s original inputs for the relationship between the propensity to respond and outcome and the response rates, but modified the model specification for generating the distribution of the $u_t$ (factors unrelated to attrition for the treatment group) and the $u_c$ (factors unrelated to attrition for the control group). The scenario was intended to roughly model the variation in the random term ($u$) in the
WWC attrition bias model. The random term contained all factors unrelated to attrition; therefore, the random error was contained in that term. Scenario 8 inserted more variation in order to roughly simulate what may happen to attrition bias estimates if a study under review used an outcome instrument with much lower reliability than another study being reviewed. As indicated in Chapter 3, even large-scale, high-stakes statewide tests may vary significantly in their reliabilities (from .61 to .89 in one Ohio example) (Ohio Department of Education, 2009).

The results from the Scenario 8 model indicated some movement upward of the bias estimates as compared to the original WWC attrition bias model (Table 4 in this study). Specifically, for the original WWC bias model, 31 percent of the 160 cells were under the .05 threshold of tolerable bias. For the Scenario 8 model, 21 percent of the 160 cells were under the .05 threshold of tolerable bias.
Table 11

*Results of Simulation Scenario 8: Simulating Sensitivity of Model to Variation in Reliability of Outcome Instrument Across Studies*

<table>
<thead>
<tr>
<th>Pt</th>
<th>Prc</th>
<th>αc=0.05</th>
<th>αc=0.05</th>
<th>αc=0.05</th>
<th>αc=0.15</th>
<th>αc=0.20</th>
<th>αc=0.20</th>
<th>αc=1.00</th>
<th>αc=−1.00</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.9</td>
<td>0.9</td>
<td>0.02</td>
<td>0.03</td>
<td>0.07</td>
<td>0.02</td>
<td>0.04</td>
<td>0.08</td>
<td>-0.01</td>
<td>0.63</td>
</tr>
<tr>
<td>0.89</td>
<td>0.91</td>
<td>0.03</td>
<td>0.04</td>
<td>0.06</td>
<td>0.03</td>
<td>0.05</td>
<td>0.10</td>
<td>0.05</td>
<td>0.64</td>
</tr>
<tr>
<td>0.875</td>
<td>0.925</td>
<td>0.04</td>
<td>0.06</td>
<td>0.08</td>
<td>0.05</td>
<td>0.08</td>
<td>0.14</td>
<td>0.11</td>
<td>0.64</td>
</tr>
<tr>
<td>0.865</td>
<td>0.935</td>
<td>0.06</td>
<td>0.07</td>
<td>0.09</td>
<td>0.08</td>
<td>0.10</td>
<td>0.16</td>
<td>0.15</td>
<td>0.62</td>
</tr>
<tr>
<td>0.85</td>
<td>0.95</td>
<td>0.07</td>
<td>0.08</td>
<td>0.11</td>
<td>0.11</td>
<td>0.13</td>
<td>0.20</td>
<td>0.20</td>
<td>0.62</td>
</tr>
<tr>
<td>0.8</td>
<td>0.8</td>
<td>0.02</td>
<td>0.04</td>
<td>0.08</td>
<td>0.04</td>
<td>0.05</td>
<td>0.13</td>
<td>0.02</td>
<td>1.00</td>
</tr>
<tr>
<td>0.79</td>
<td>0.81</td>
<td>0.03</td>
<td>0.05</td>
<td>0.09</td>
<td>0.04</td>
<td>0.06</td>
<td>0.16</td>
<td>0.03</td>
<td>1.01</td>
</tr>
<tr>
<td>0.775</td>
<td>0.825</td>
<td>0.04</td>
<td>0.07</td>
<td>0.09</td>
<td>0.06</td>
<td>0.10</td>
<td>0.17</td>
<td>0.09</td>
<td>0.98</td>
</tr>
<tr>
<td>0.765</td>
<td>0.835</td>
<td>0.05</td>
<td>0.08</td>
<td>0.12</td>
<td>0.08</td>
<td>0.12</td>
<td>0.20</td>
<td>0.11</td>
<td>1.00</td>
</tr>
<tr>
<td>0.75</td>
<td>0.85</td>
<td>0.07</td>
<td>0.09</td>
<td>0.13</td>
<td>0.10</td>
<td>0.13</td>
<td>0.22</td>
<td>0.16</td>
<td>0.96</td>
</tr>
<tr>
<td>0.7</td>
<td>0.7</td>
<td>0.03</td>
<td>0.06</td>
<td>0.11</td>
<td>0.04</td>
<td>0.06</td>
<td>0.17</td>
<td>0.01</td>
<td>1.28</td>
</tr>
<tr>
<td>0.69</td>
<td>0.71</td>
<td>0.04</td>
<td>0.08</td>
<td>0.12</td>
<td>0.06</td>
<td>0.07</td>
<td>0.19</td>
<td>0.03</td>
<td>1.29</td>
</tr>
<tr>
<td>0.675</td>
<td>0.725</td>
<td>0.06</td>
<td>0.08</td>
<td>0.14</td>
<td>0.08</td>
<td>0.11</td>
<td>0.22</td>
<td>0.08</td>
<td>1.28</td>
</tr>
<tr>
<td>0.665</td>
<td>0.735</td>
<td>0.05</td>
<td>0.10</td>
<td>0.13</td>
<td>0.09</td>
<td>0.12</td>
<td>0.23</td>
<td>0.12</td>
<td>1.33</td>
</tr>
<tr>
<td>0.65</td>
<td>0.75</td>
<td>0.07</td>
<td>0.09</td>
<td>0.15</td>
<td>0.10</td>
<td>0.14</td>
<td>0.26</td>
<td>0.16</td>
<td>1.30</td>
</tr>
<tr>
<td>0.6</td>
<td>0.6</td>
<td>0.03</td>
<td>0.07</td>
<td>0.13</td>
<td>0.05</td>
<td>0.08</td>
<td>0.21</td>
<td>0.00</td>
<td>1.60</td>
</tr>
<tr>
<td>0.59</td>
<td>0.61</td>
<td>0.06</td>
<td>0.09</td>
<td>0.14</td>
<td>0.06</td>
<td>0.09</td>
<td>0.23</td>
<td>0.03</td>
<td>1.60</td>
</tr>
<tr>
<td>0.575</td>
<td>0.625</td>
<td>0.07</td>
<td>0.09</td>
<td>0.16</td>
<td>0.08</td>
<td>0.12</td>
<td>0.25</td>
<td>0.07</td>
<td>1.60</td>
</tr>
<tr>
<td>0.565</td>
<td>0.635</td>
<td>0.07</td>
<td>0.10</td>
<td>0.17</td>
<td>0.09</td>
<td>0.14</td>
<td>0.26</td>
<td>0.11</td>
<td>1.58</td>
</tr>
<tr>
<td>0.55</td>
<td>0.65</td>
<td>0.08</td>
<td>0.12</td>
<td>0.18</td>
<td>0.12</td>
<td>0.16</td>
<td>0.30</td>
<td>0.16</td>
<td>1.59</td>
</tr>
</tbody>
</table>

Green cells meet the .05 bias threshold
Chapter 5: Discussion, Conclusions, and Recommendations

This dissertation focused on replicating and extending the U.S. Department of Education What Work Clearinghouse model of attrition bias that was developed and continues to be used as one of the main arbiters of the research that gets reviewed and accepted into the WWC. The WWC was developed specifically to enhance the rigor of educational research and provide educators a source for quality research on practices focused on improving student achievement. As Slavin (2008) stated, “Potentially, the WWC is the most important of the synthesis efforts for policy, because it alone carries the endorsement of the U.S. Department of Education.” (p. 6)

The work of the WWC-commissioned researchers to examine attrition bias systematically should be lauded. The WWC was given a critical charge—develop a way to separate good research that can inform practice from poorly-executed research where findings may even do more harm than good for children and families. Before the attrition guidelines were established empirically by the WWC, there was little guidance on how much attrition is too much attrition and arbitrary thresholds were set in many fields of inquiry, including education (Amico, 2009; Valentine and McHugh, 2007). The guidelines set by the WWC in its procedures and standards handbook are now being used not only for educational research, but by other fields such as child welfare (U.S. Department of Health and Human Services, 2013).

For education researchers, the WWC-commissioned attrition guidelines can determine how a study is rated, thereby a judgment of quality and utility of the findings. If a study is accepted for review by WWC researchers, reviewers use a rigorous set of
criteria to determine if the study will be rated as “meets evidence standards,” “meets evidence standards with reservations,” or “does not meet evidence standards.” One of the key review criteria is the amount and type of attrition. If the study has some combination of overall and differential attrition that indicates high levels of potential bias (recall the blue zone in Figure 1), it can be rated no higher than “meets evidence standards with reservations” and, if reviewers are not provided with evidence that there was baseline equivalence in the treatment and control groups, the study would likely be rejected for inclusion in the WWC. If a study falls in the yellow zone, it is at the discretion of the reviewers to weigh the evidence and decide if the attrition is largely exogenous or endogenous. If a study falls in the green zone, attrition will not prohibit the study from being labeled as “meets evidence standards” if the study also holds up to the other criteria in the review process such as attribution of effect or equivalence of the intervention and comparison groups at baseline (U.S. Department of Education, 2008, 2011b).

If the study falls in the yellow zone of Figure 1 based on the overall and differential attrition reported, WWC reviewers are then tasked with making a decision about how much they think the attrition is exogenous—unrelated to the outcome, or endogenous—related to the outcome. They are then charged with being either “optimistic” about the relationship between attrition and outcome, or “conservative” about the relationship between attrition and outcome (U.S. Department of Education, 2011, p. 35). The reviewers have an important task and a fair amount of discretion in determining whether the reported attrition will label a large-scale educational study as not meeting evidence standards.
Clearly, the attrition bias thresholds illustrated in Figure 1 and used by WWC reviewers are consequential. They were developed by WWC researchers who first developed a model of attrition bias, then set the bias thresholds via a Monte Carlo study that varied rates of overall and differential attrition and modeled eight different scenarios of how much a participant’s the propensity to respond is related to his or her outcome. This dissertation replicated the WWC scenarios and extended the work by developing and simulating additional scenarios in order to test the sensitivity of the WWC model and provides additional information regarding the bias thresholds. This chapter includes a discussion of the results of the simulation scenarios addressing the research questions. Conclusions regarding the sensitivity analyses are also included, as well as recommendations for further research.

**Discussion of the Results**

The first research question addressed in this dissertation was: *How do the amount and type of attrition, under varying assumptions of how much subjects’ likelihood of dropping out is related to their outcome (MNAR), impact randomized controlled studies by contributing to bias, based on the What Works Clearinghouse model of bias?* Addressing this research question required independent replication of the WWC Monte Carlo simulation based on their model of attrition bias. This model was replicated successfully as indicated in Table 4. This original WWC model indicates that varying the amount and type of attrition (overall and differential) as well as the relationship between the subjects’ likelihood of dropping out and their outcome, contributes to bias in RCTs.
The WWC bias estimates, as evident in the heat table color patterns illustrated in Table 4, are affected by the interaction of attrition and the relationship between the propensity to respond and outcome. For example, Column 1 of Tables 2 and 4 is the scenario with the lowest values of the relationship between the propensity to respond and outcome (.075 for the treatment group and .05 for the control group). Approximately 80 percent of the bias estimates in Column 1 fall below the .05 bias threshold. In other words, the overall and differential attrition do not inflate the bias estimate in the situation where the relationship between the propensity to respond and outcome is considered weak and not very different between the treatment and control group. As the relationship between the propensity to respond and outcome gradually increases in magnitude (Columns 2 through 8), the bias estimates get larger. While this pattern holds true across the table, another pattern is evident. Columns 1, 4, and 7, where the difference between the treatment and control group in the relationship of the propensity to respond and outcome is not that large, have the most cells with bias estimates below the threshold. The pattern indicates that the difference in the relationship between the propensity to respond and outcome in the original WWC model are as important to the bias estimates as the overall and differential attrition.

The second research question was: *How sensitive is the What Works Clearinghouse model of bias to changes in attrition rates and/or the relationship between outcome and a participant’s propensity to respond in randomized controlled trials?* This research question was addressed by modeling scenarios 2 through 7 as presented in Chapter 3. Scenario 2 confirmed the model specification by eliminating all relationship
between the propensity to respond and outcome. This modeled a situation where no outcome data were MNAR. As expected, the bias estimates were held at zero or near zero regardless of the amount and type of attrition.

Scenarios 3 and 4 use the same overall and differential attrition rates as the WWC study, but add 23 additional unique variations in the relationship between the propensity to respond and outcome for the treatment and control group. Across all 23 variations, the columns where that difference between the relationship of attrition and outcome for the treatment and control group (the difference between $\alpha_t$ and $\alpha_c$) is held to as low as .05 have more bias estimates under the threshold of acceptable bias. This even holds true when the relationship between the propensity to respond and outcome is at .35 for the treatment group and .30 for the control group (Column 15 on Table 7). In that column, 45 percent of the cells have acceptable bias estimates, even though the relationship between a participant’s propensity to respond and outcome is large and the overall and differential attrition rates vary. This portion of the sensitivity analysis illustrates the importance of the assumptions made about how much attrition is related to outcome in the treatment versus the control group.

Another pattern emerged from the Scenario 3 and Scenario 4 results. As illustrated by the stepped pattern in the heat colors in Tables 6 and 7, when differential attrition increases along with the magnitude of difference in the propensity to respond and outcome between the treatment and control groups, bias estimates get larger. This indicates the sensitivity of the model to the interaction between the differential attrition
rate and the assumption about the relationship between the propensity to respond and outcome.

Scenario 5 extended the question about the impact of the difference in magnitude of the relationship between attrition and outcome by setting the control group relationship between the propensity to respond and outcome to zero. The WWC model was very sensitive to this modification, as only one cell indicated acceptable bias. Even with a relatively low relationship of the propensity to respond to outcome for the treatment group (.075) in Column 1 of Table 8, all but one cell was outside the acceptable bias range. The same results held true for Scenario 6, where the control group relationship between the propensity to respond and outcome was set to zero and the attrition in the control group was held at 10 percent.

Scenario 7 modeled the sensitivity of the WWC model to overall and differential attrition rates. In this scenario, the relationship between the propensity to respond and outcome was set as equal and increased only in magnitude from .075 to 1.0 for both the treatment and control groups. As the Table 10 heat colors illustrate, there is a clear, stepped pattern to the bias estimates as the differential attrition rates increase as the overall relationship between the propensity to respond and outcome increases.

The third research question was: How does varying the random error to simulate variations in the reliability of instruments used across studies impact the attrition thresholds? Because the WWC researchers modeled the portion of the outcome that is unrelated to attrition as one random term (the \( u \) term in the formulas), that random term by definition must include any random error introduced by measurement error or based
on the reliability of the outcome instrument itself. Since the reliability of the outcome instruments used varies across studies that are reviewed by the WWC, attempting to gauge the sensitivity of the attrition bias model to changes in the magnitude of that random term gives some indication of the potential impact of reliability on the attrition bias thresholds.

This dissertation modeled the potential impact of variations in instrument reliability across studies by increasing the standard deviation of the distribution used to model the portion of the outcome that is unrelated to attrition. The results, as presented in Table 11, indicate that the variation in reliability of the outcome instrument may impact the bias thresholds. Holding everything else constant in the original WWC attrition bias model, increasing the standard deviation of the error term from 1 to 1.2 resulted in 10 percent fewer cells with bias estimates under the .05 threshold.

The fourth research question was: *Are there discernible patterns to the increase or decrease in bias at different increments of overall and differential attrition and propensity to respond in randomized controlled trials?* The results indicate that there are discernible patterns in the bias estimates when the increments of overall and differential attrition and/or the relationship of the propensity to respond and outcome, or propensity to respond, are adjusted. These were illustrated in the stepped patterns of heat colors, most notably in Tables 6, 7, and 10. As the relationship between the propensity to respond and outcome becomes more divergent between the treatment and control group, the bias estimates get larger in a discernible pattern. Also, as the magnitude of differential
attrition increases in conjunction with the increase in the relationship between the propensity to respond and outcome, bias estimates get larger.

**Conclusions**

“Quis custodiet ipsos custiodes” or “Who will guard the guards themselves?”—a phrase from the Roman poet Juvenal—seems appropriate to summarize the purpose of this study (Green, 1982). The U.S. Department of Education developed the WWC to be the guardian of high quality research on “what works” in education. This task is consequential to children and families, as educators need to be able to trust the research on interventions that purport to positively impact student achievement. Knowing the best interventions for individual students is more than an educational issue, but a social justice issue for poor and minority children.

The first task of this study was to independently replicate the WWC-commissioned researchers’ Monte Carlo simulation that developed the current attrition bias thresholds. This in itself was challenging, as the WWC researchers only provided the bias formula and few details about how the simulation was programmed. (It should be noted that it was not the purpose of the researchers to provide these details in the WWC Procedures and Standards Handbook. The researchers were not likely charged with developing a scholarly publication, but a set of practical guidelines for other researchers intending to submit their work for review and inclusion in the WWC.) However, this work’s independent and nearly exact replication of the WWC’s original study provides confirmation of the results and the soundness of the methodology—a valuable finding in itself.
In addition to successfully replicating the Monte Carlo study that was used to set the WWC attrition guidelines, this dissertation primarily addressed the sensitivity of the WWC model of bias to changes in attrition rates and/or the relationship between outcome and a participant’s propensity to respond in RCTs. Changes in assumptions about the size of the random term in the model were also utilized to address the sensitivity of the model to variability in the reliability of the outcome instrument from one study to another.

The dissertation was able to validate the original study’s attrition bias thresholds and the sensitivity analyses indicate that the WWC attrition bias model is sensitive to changes in the assumptions about the relationship between attrition and outcome. That sensitivity is perhaps the most important finding of this dissertation that was not evident in the original WWC table of bias estimates. As the What Works Clearinghouse Procedures and Standards Handbook states,

The PI considers the types of samples and likely relationship between attrition and student outcomes for studies in the topic area. In cases where the PI has reason to believe that much of the attrition is exogenous—such as parent mobility with young children—more optimistic assumptions regarding the relationship between attrition and the outcome might be appropriate. On the other hand, in cases such as high school students choosing whether to participate in an intervention—more conservative assumptions may be appropriate. (U.S. Department of Education, 2011b, p. 14)

This responsibility on the PI (now referred to as review team leader) is a consequential one, as under the attrition bias model, any study that falls in the middle region (yellow in
Figure 1) cannot be labeled as meeting evidence standards unless the reviewer assumes that there is little relationship between the propensity to respond and outcome. The WWC Procedures and Standards Handbook clearly states that it is up to the reviewer’s judgment to determine how much he or she thinks the attrition is related to the outcome.

Figure 1 was developed from the bias estimates produced by the WWC researchers and provided in Table 2 of this study, then replicated in Table 4. What these dissertation findings indicate is that the bias estimates are very sensitive to small changes in the assumptions about the relationship of the propensity to respond to outcome. Moving from a scenario with a differential of \( \alpha_t = .25, \alpha_c = .15 \) (Table 7, Column 3) to a scenario with a differential of \( \alpha_t = .25, \alpha_c = .20 \) (Table 7, Column 4) more than doubles the number of bias estimates that fall below the acceptable threshold. However, the difference between an \( \alpha_c \) of .15 to one of .20 is not a large one. Also, as the WWC indicates and as these dissertation results confirm, if the relationship between the propensity to respond and outcome at follow up is the same for both the treatment and the control groups, then attrition bias is only an issue when the differential attrition gets larger (Table 10).

These results raise the question of how accurate the reviewers need to be in judging how much the propensity to respond is related to the outcome. If the reviewers assume what the WWC calls the most optimistic assumption of the relationship of the propensity to respond to outcome, or \( \alpha_t \) at .075 and \( \alpha_c \) at .05, how can the reviewer be certain within a tolerance of .025 that the relationship between the propensity to respond and outcome is not the same (0.075 or .05) in both treatment and control groups, which
according to the WWC attrition bias model and illustrated in Table 10 of this dissertation yields no bias, regardless of the amount of overall and differential attrition and thereby rendering the regions in Figure 1 too conservative? As Bell and colleagues found in their 2013 study on differential attrition and bias in RCTs, equal dropout rates between treatment and control participants do not imply that the estimates of treatment effect are unbiased, and unequal dropout rates do not imply that the estimates of treatment effect are necessarily biased (Bell et al., 2013).

The conclusion reached by varying the standard deviation in the distribution containing the random error term in the model in order to address the potential impact of reliability on the bias thresholds indicates that the WWC attrition bias thresholds may be somewhat sensitive to varying reliabilities of instruments across studies. The WWC guidance for reviewers indicated that the outcome measure must be valid and reliable, but within any acceptable reliability threshold (not stated in the guidance) there will be variation in reliability across studies. This sensitivity may call for more reviewer guidance, especially for studies that fall into the yellow zone of Figure 1 where reviewer discretion determines their classification of meeting evidence standards.

**Limitations**

This research did not include information from those who have reviewed RCTs for inclusion into the WWC. Information on any additional, specific guidance they received (beyond the guidance in the WWC Procedures and Standards Handbook) during their reviewer trainings on how to make assumptions or judgments about the relationship between the propensity to respond and outcome would add value to this study’s findings.
This dissertation replicated and extended the original WWC attrition bias model and simulation by adding additional scenarios of attrition and magnitude of outcome data MNAR. However, the study did not modify the formulas to provide alternative models for estimating bias. For example, since this study further illuminated the sensitivity of the model to the difference between the treatment group and control group in the magnitude of relationship of attrition and outcome. This sensitivity was best illustrated in Tables 6 and 7, where the differences between the treatment and control groups in the relationship between the propensity to respond and outcome tended to override the differences in the magnitude of overall and differential attrition. Changing the attrition bias formula to allow for the relationship between the propensity to respond and outcome to be less impactful in the model until that relationship reached a certain threshold of overall magnitude may help provide some more specific guidance to reviewers who may be reviewing studies where they assume zero or near zero relationship of the propensity to respond to outcome in the control group. As the model indicates now, assuming a zero relationship of the propensity to respond to outcome in the control group and a relatively low relationship between the propensity to respond to outcome in the treatment group (i.e. 0.075) results in unacceptable levels of bias regardless of the rate of overall or differential attrition.

Recommendations for Future Research

A meta-analysis of the studies that have been reviewed by the WWC since its inception, examining the reported overall and differential attrition rates and any data that provide insight into the relationship between the propensity to respond and outcome (e.g.
the full sample demographics and the respondent sample demographics), could serve to
guide refinements to the attrition bias thresholds and guidance for reviewers. As this
study further illuminated, differential attrition can be very high as long as the difference
between the treatment and control group in the relationship of the propensity to respond
and outcome is small. A meta-analysis may help indicate where the attrition bias
guidelines excluded some studies where researchers could plausibly assume little
difference in the relationship between the propensity to respond and outcome in the
treatment versus the control group.

The same meta-analysis could examine the reported reliabilities of the outcome
instruments used across the studies that have been reviewed by the WWC may illuminate
the variation or lack of variation in reliability. If there is little variation and all studies
that have been accepted for review have relatively high reliabilities, then no further work
to examine the relationship of instrument reliability to the WWC bias model is warranted.
If there is large variation in the reliabilities, then it is important to gauge how the existing
model may be including that or needs to be augmented in order to include it.

Modifying the attrition bias formula to allow for the relationship between the
propensity to respond and outcome to be less impactful in the model until that
relationship reaches a certain threshold overall magnitude may help provide some more
specific guidance to reviewers who may be reviewing studies where they assume zero or
near zero relationship of the propensity to respond to outcome in the control group.
Recommendations for Educational Researchers and Reviewers of Educational Research

The findings of the original WWC-commissioned researchers, and the extension of those findings based on this study, regarding the consequences of different rates of overall and differential attrition and/or the extent to which a participant’s propensity to respond to the outcome measure is correlated with his or her outcome itself is a clear message to those reviewing research for the WWC or other entities or fields. As indicated in this study’s findings, the decision made by the reviewer regarding the extent to which the propensity to respond is related to the outcome is critical, as this decision in itself may determine if a study meets or does not meet evidence standards. This dissertation illustrated the sensitivity of the attrition bias model to relatively small changes in the differential between the propensity to respond/outcome relationship in the treatment group vs. the control group. Perhaps reviewers of studies such as those put before the WWC should be required to document and make a case their decisions about how much the propensity to respond and outcome are likely related.

This dissertation’s extension of the original study to model the additive effects on bias based on the reliability of the outcome instrument calls for more attention by both the researchers conducting the studies and the reviewers on the quality of the instruments deployed. The WWC Procedures and Standards Handbook Version 3.0 (currently under review but not yet the official handbook for researchers) sets reliability thresholds at 0.50 for internal consistency reliability, 0.40 for test-retest reliability (U.S. Department of Education, 2013). Therefore, the WWC reviewers may be applying the same attrition bias
thresholds to one study with a 0.50 internal consistency reliability and another with a 0.89 internal consistency reliability. The findings in this dissertation indicate that varying reliabilities have a consequential impact on the attrition bias thresholds. Reviewers need to require high instrument reliabilities and researchers need to focus on the quality of the instruments used in order to guard against additional bias inserted by poorer-quality instruments.

High-quality, large-scale research on what works best in the classroom to raise student achievement, the work that drives the purpose of the WWC, has only grown in importance since the WWC was formed. Race to the Top, the largest infusion of federal dollars for educational improvement in history under a competitive grant process, requires that teacher performance evaluations now include a measure of students’ academic growth (United States Government Accountability Office, 2013; Wood, Moore, Clarkwest, Killewald, and Monahan, 2012). Evidence-based practices to improve student achievement are high stakes at the individual teacher and principal level now rather than just at the district or building level. Rigorous methodology that holds the bar high for educational research, yet does not arbitrarily exclude good work that could inform practice, is critical to the field. This rigorous methodology includes the WWC attrition guidelines.

The attrition guidelines originally designed for experimental and quasi-experimental studies under review by the WWC are now being applied to studies using other methodologies such as single-case designs (U.S. Department of Education, 2011b). In addition, other federal agencies are adopting similar review guidelines and at least one,
the U.S. Department of Health and Human Services, adopted the WWC attrition standards as part of their evidence review protocol for rating research on families and children (U.S. Department of Health and Human Services, 2013). Continuing to refine those guidelines and to examine how they impact the WWC’s utility is important work for education and other fields.
References


simulation study. Unpublished manuscript, Peabody Research Institute, Vanderbilt University, Nashville, TN.


Appendix A: Scenario 1 R Code Replicating Table A1 from What Works
Clearinghouse Procedures and Standards Handbook

#R version 3.0.2 (2013-09-25) -- "Frisbee Sailing"
#Copyright (C) 2013 The R Foundation for Statistical Computing
#Platform: x86_64-w64-mingw32/x64 (64-bit)

rm(list = ls())

#setting seed for random number generator so results can be replicated exactly
set.seed(18040817)

#50,000 replications
n <- 50000

#inputting vectors of WWC values for proportion of variation of outcome explained by the propensity to respond
#treatment group:
a1 <- c(0.075, 0.1, 0.15, 0.2, 0.3, 0.5, 1, 1)
#control group
a2 <- c(0.05, 0.05, 0.05, 0.15, 0.2, 0.2, 1, 1)

numalpha <- length(a1)

#creating matrix from vectors created above (proportion of variation of outcome explained by the propensity to respond)
a <- matrix(c(a1,a2), nrow=2, ncol=numalpha, byrow=TRUE)

#taking square root of proportion of variation of the propensity to respond explained by outcome. WWC refers to alpha #as similar to R², so square root is the correlation
a <- sqrt(a)
a

#delineating negative 1 for row 8 in control group from WWC Appendix A, Table A1
a[2,8] <- a[2,8]*-1
a

#inputting vectors of WWC values for response rates in
treatment and control groups
p1 <- c(.9, .89, .875, .865, .85, .8, .79, .775, .765, .75, .7, .69, .675, .665, .65, .6, .59, .575, .565, .55)
p2 <- c(.9, .91, .925, .935, .95, .8, .81, .825, .835, .85, .7, .71, .725, .735, .75, .6, .61, .625, .635, .65)
numprobs <- length(p1)

#creating matrix from vectors of response rates
pp  <- matrix(c(p1,p2), nrow=numprobs, ncol=2, byrow=FALSE)

#creating formula for threshold of propensity to respond
p <- 1-pp

#calling quantile function (if z is greater than the value
that corresponds to a percentile of the z distribution
(given ρ), then individual responds at follow-up
q <- qnorm(p)

b   <- matrix(data=NA,nrow=numprobs,ncol=numalpha)

nYt <- matrix(data=NA,nrow=numprobs,ncol=numalpha)
pYt <- matrix(data=NA,nrow=numprobs,ncol=numalpha)

nYc <- matrix(data=NA,nrow=numprobs,ncol=numalpha)
pYc <- matrix(data=NA,nrow=numprobs,ncol=numalpha)

mYt <- matrix(data=NA,nrow=numprobs,ncol=numalpha)
mYc <- matrix(data=NA,nrow=numprobs,ncol=numalpha)

maxi <- as.integer(numprobs)
maxj <- as.integer(numalpha)
mini <- 1
maxi <- maxi
minj <- 1
maxj <- maxj
for (i in mini:maxi)
{
  for (j in minj:maxj)
  {
    Yt <- matrix(data=NA,nrow=n,ncol=1,byrow=FALSE)
    Yc <- matrix(data=NA,nrow=n,ncol=1,byrow=FALSE)
#creating cumulative normal distributions for z and u
(z=individual’s propensity to respond; u=factor unrelated to attrition [random error])

Zt <- rnorm(n)
Zc <- rnorm(n)
Ut <- rnorm(n)
Uc <- rnorm(n)

#inputting the WWC formula for outcome at follow up
for (k in 1:n)
{
  if (Zt[k] > q[i,1]) {Yt[k] <- (a[1,j] * Zt[k]) + (1-a[1,j]) * Ut[k]}
  if (Zc[k] > q[i,2]) {Yc[k] <- (a[2,j] * Zc[k]) + (1-a[2,j]) * Uc[k]}
}

nYt[i,j] <- length(which(!is.na(Yt)))
nYc[i,j] <- length(which(!is.na(Yc)))
pYt[i,j] <- nYt[i,j]/n
pYc[i,j] <- nYc[i,j]/n
mYt[i,j] <- mean(Yt, na.rm=TRUE)
mYc[i,j] <- mean(Yc, na.rm=TRUE)
#b[i,j] <- mYt[i,j]-mYc[i,j]
}

#WWC bias formula
gb <- (mYt - mYc)
gb <- round(gb,3)
gb
Appendix B: Scenario 2 R Code Modeling the Counterfactual

#R version 3.0.2 (2013-09-25) -- "Frisbee Sailing"
#Copyright (C) 2013 The R Foundation for Statistical Computing
#Platform: x86_64-w64-mingw32/x64 (64-bit)

rm(list = ls())

#setting seed for random number generator so results can be replicated exactly
set.seed(18040817)

#50,000 replications
n <- 50000

#inputting vectors of WWC values for proportion of variation of outcome explained by the propensity to respond
treatment group:
a1 <- c(0)
#control group
a2 <- c(0)

numalpha <- length(a1)
numalpha

#creating matrix from vectors created above (proportion of variation of outcome explained by the propensity to respond)
a <- matrix(c(a1,a2), nrow=2, ncol=numalpha, byrow=TRUE)

#taking square root of proportion of variation of the propensity to respond explained by outcome. WWC refers to alpha #as similar to R^2, so square root is the correlation
a <- sqrt(a)
a

#inputting vectors of WWC values for response rates in treatment and control groups
p1 <- c(1,.9, .89, .875, .865, .85, .8, .79, .775, .765, .75, .7, .69, .675, .665, .65, .6, .59, .575, .565, .55)
p2 <- c(1, .9, .91, .925, .935, .95, .8, .81, .825, .835, .85, .7, .71, .725, .735, .75, .6, .61, .625, .635, .65)
numprobs <- length(p1)
numprobs

# creating matrix from vectors of response rates
pp <- matrix(c(p1, p2), nrow=numprobs, ncol=2, byrow=FALSE)

# creating formula for threshold of propensity to respond
p <- 1-pp

# calling quantile function (if z is greater than the value that corresponds to a percentile of the z distribution (given \( \rho \)), then individual responds at follow-up
q <- qnorm(p)

b <- matrix(data=NA, nrow=numprobs, ncol=numalpha)

nYt <- matrix(data=NA, nrow=numprobs, ncol=numalpha)
pYt <- matrix(data=NA, nrow=numprobs, ncol=numalpha)
nYc <- matrix(data=NA, nrow=numprobs, ncol=numalpha)
pYc <- matrix(data=NA, nrow=numprobs, ncol=numalpha)
mYt <- matrix(data=NA, nrow=numprobs, ncol=numalpha)
mYc <- matrix(data=NA, nrow=numprobs, ncol=numalpha)

maxi <- as.integer(numprobs)
maxj <- as.integer(numalpha)
mini <- 1
maxi <- maxi
minj <- 1
maxj <- maxj
for (i in mini:maxi)
{
  for (j in minj:maxj)
  {
    Yt <- matrix(data=NA, nrow=n, ncol=1, byrow=FALSE)
    Yc <- matrix(data=NA, nrow=n, ncol=1, byrow=FALSE)

    # creating cumulative normal distributions for z and u (z=individual’s propensity to respond; u=factor unrelated to attrition [random error])
    Zt <- rnorm(n)
    Zc <- rnorm(n)
Ut <- rnorm(n)
Uc <- rnorm(n)

# inputting the WWC formula for outcome at follow up
for (k in 1:n)
{
  if (Zt[k] > q[i,1]) {Yt[k] <- (a[1,j] * Zt[k]) + (1-a[1,j]) * Ut[k])}
  if (Zc[k] > q[i,2]) {Yc[k] <- (a[2,j] * Zc[k]) + (1-a[2,j]) * Uc[k])}
}
nYt[i,j] <- length(which(!is.na(Yt)))
nYc[i,j] <- length(which(!is.na(Yc)))
pYt[i,j] <- nYt[i,j]/n
pYc[i,j] <- nYc[i,j]/n
mYt[i,j] <- mean(Yt, na.rm=TRUE)
mYc[i,j] <- mean(Yc, na.rm=TRUE)
#b[i,j] <- mYt[i,j]-mYc[i,j]
}
}

# WWC bias formula
gb <- (mYt - mYc)
gb <- round(gb,3)
gb
Appendix C: Scenario 3 R Code Systematically Extending Alphas and Response Rates for Sensitivity Analysis

#R version 3.0.2 (2013-09-25) -- "Frisbee Sailing"
#Copyright (C) 2013 The R Foundation for Statistical Computing
#Platform: x86_64-w64-mingw32/x64 (64-bit)

rm(list = ls())

#setting seed for random number generator so results can be replicated exactly
set.seed(18040817)

#50,000 replications
n <- 50000

#inputting vectors of EXTENDED values (SET 1) for proportion of variation of outcome explained by the propensity to respond
#treatment group:
a1 <- c(.075, .075,.085,.085,.095,.095,.10,.10,.15,.15,.20,.20,.20)
#control group
a2 <- c(.05,.10,.05,.05,.05,.05,.05,.05,.10,.05,.10,.05,.10,.15)

numalpha <- length(a1)
numalpha

#creating matrix from vectors created above (proportion of variation of outcome explained by the propensity to respond)
a <- matrix(c(a1,a2), nrow=2, ncol=numalpha, byrow=TRUE)

#taking square root of proportion of variation of the propensity to respond explained by outcome. WWC refers to alpha as similar to $R^2$, so square root is the correlation
a <- sqrt(a)
a
# inputting vectors of values for response rates in
treatment and control groups
p1 <- c(.9, .89, .875, .865, .85, .8, .79, .775, .765, .75,
    .7, .69, .675, .665, .65, .6, .59, .575, .565, .55)
p2 <- c(.9, .91, .925, .935, .95, .8, .81, .825, .835, .85,
    .7, .71, .725, .735, .75, .6, .61, .625, .635, .65)
numprobs <- length(p1)

# creating matrix from vectors of response rates
pp <- matrix(c(p1, p2), nrow=numprobs, ncol=2, byrow=FALSE)

# creating formula for threshold of propensity to respond
p <- 1-pp

# calling quantile function (if \( z \) is greater than the value
# that corresponds to a percentile of the \( z \) distribution
# (given \( \rho \)), then individual responds at follow-up
q <- qnorm(p)

b <- matrix(data=NA, nrow=numprobs, ncol=numalpha)
Yt <- matrix(data=NA, nrow=n, ncol=1, byrow=FALSE)
Yc <- matrix(data=NA, nrow=n, ncol=1, byrow=FALSE)
#creating cumulative normal distributions for z and u
(z=individual’s propensity to respond; u=factor unrelated to attrition [random error])
Zt <- rnorm(n)
Zc <- rnorm(n)
Ut <- rnorm(n)
Uc <- rnorm(n)

#inputting the WWC formula for outcome at follow up
for (k in 1:n)
{
  if (Zt[k] > q[i,1]) {Yt[k] <- (a[1,j] * Zt[k]) + (1-a[1,j]) * Ut[k]}
  if (Zc[k] > q[i,2]) {Yc[k] <- (a[2,j] * Zc[k]) + (1-a[2,j]) * Uc[k]}
}
nYt[i,j] <- length(which(!is.na(Yt)))
nYc[i,j] <- length(which(!is.na(Yc)))
pYt[i,j] <- nYt[i,j]/n
pYc[i,j] <- nYc[i,j]/n
mYt[i,j] <- mean(Yt, na.rm=TRUE)
mYc[i,j] <- mean(Yc, na.rm=TRUE)
#b[i,j] <- mYt[i,j]-mYc[i,j]
}

#WWC bias formula
gb <- (mYt - mYc)
gb <- round(gb,3)
gb
Appendix D: Scenario 4 R Code Systematically Extending Alphas and Response

Rates for Sensitivity Analysis, Part 2

#R version 3.0.2 (2013-09-25) -- "Frisbee Sailing"
#Copyright (C) 2013 The R Foundation for Statistical Computing
#Platform: x86_64-w64-mingw32/x64 (64-bit)

rm(list = ls())

#setting seed for random number generator so results can be replicated exactly
set.seed(18040817)

#50,000 replications
n <- 50000

#inputting vectors of WWC values for proportion of variation of outcome explained by response
#treatment group:
#control group
a2 <- c(.05,.10,.15,.20,.05,.10,.15,.20,.25,.05,.10,.15,.20,.25,.30)

numalpha <- length(a1)
numalpha

#creating matrix from vectors created above (proportion of variation of outcome explained by the propensity to respond)
a <- matrix(c(a1,a2), nrow=2, ncol=numalpha, byrow=TRUE)

#taking square root of proportion of variation of the propensity to respond explained by outcome. WWC refers to alpha #as similar to R^2, so square root is the correlation
a <- sqrt(a)
a
#inputting vectors of WWC values for response rates in treatment and control groups
p1 <- c(.9, .89, .875, .865, .85, .8, .79, .775, .765, .75, .7, .69, .675, .665, .65, .6, .59, .575, .565, .55)
p2 <- c(.9, .91, .925, .935, .95, .8, .81, .825, .835, .85, .7, .71, .725, .735, .75, .6, .61, .625, .635, .65)
numprobs <- length(p1)
numprobs

#creating matrix from vectors of response rates
pp <- matrix(c(p1,p2), nrow=numprobs, ncol=2, byrow=FALSE)

#creating formula for threshold of propensity to respond
p <- 1-pp

#calling quantile function (if z is greater than the value that corresponds to a percentile of the z distribution (given \( \rho \)), then individual responds at follow-up
q <- qnorm(p)

b <- matrix(data=NA,nrow=numprobs,ncol=numalpha)
nYt <- matrix(data=NA,nrow=numprobs,ncol=numalpha)
pYt <- matrix(data=NA,nrow=numprobs,ncol=numalpha)
nYc <- matrix(data=NA,nrow=numprobs,ncol=numalpha)
pYc <- matrix(data=NA,nrow=numprobs,ncol=numalpha)
mYt <- matrix(data=NA,nrow=numprobs,ncol=numalpha)
mYc <- matrix(data=NA,nrow=numprobs,ncol=numalpha)

maxi <- as.integer(numprobs)
maxj <- as.integer(numalpha)
mini <- 1
maxi <- maxi
minj <- 1
maxj <- maxj
for (i in mini:maxi)
{
  for (j in minj:maxj)
  {
    Yt <- matrix(data=NA,nrow=n,ncol=1,byrow=FALSE)
    Yc <- matrix(data=NA,nrow=n,ncol=1,byrow=FALSE)
# creating cumulative normal distributions for z and u  
(z=individual’s propensity to respond; u=factor unrelated to attrition [random error])  
Zt <- rnorm(n)  
Zc <- rnorm(n)  
Ut <- rnorm(n)  
Uc <- rnorm(n)  

# inputting the WWC formula for outcome at follow up  
for (k in 1:n)  
  {  
  if (Zt[k] > q[i,1]) {Yt[k] <- (a[1,j] * Zt[k]) + (1-a[1,j]) * Ut[k]}  
  if (Zc[k] > q[i,2]) {Yc[k] <- (a[2,j] * Zc[k]) + (1-a[2,j]) * Uc[k]}  
  }  
nYt[i,j] <- length(which(!is.na(Yt)))  
nYc[i,j] <- length(which(!is.na(Yc)))  
pYt[i,j] <- nYt[i,j]/n  
pYc[i,j] <- nYc[i,j]/n  
mYt[i,j] <- mean(Yt, na.rm=TRUE)  
mYc[i,j] <- mean(Yc, na.rm=TRUE)  
# b[i,j] <- mYt[i,j]-mYc[i,j]  
}  

# WWC bias formula  
gb <- (mYt - mYc)  
gb <- round(gb,3)  
gb
Appendix E: Scenario 5 R Code Setting Control Group Relationship of The propensity to respond to Outcome at Zero

```r
#R version 3.0.2 (2013-09-25) -- "Frisbee Sailing"
#Copyright (C) 2013 The R Foundation for Statistical Computing
#Platform: x86_64-w64-mingw32/x64 (64-bit)

rm(list = ls())

#setting seed for random number generator so results can be replicated exactly
set.seed(18040817)

#50,000 replications
n <- 50000

#inputting vectors of WWC values for proportion of variation of outcome explained by response
#treatment group:
a1 <- c(0.075, 0.1, 0.15, 0.2, 0.3, 0.5, 1)
#control group
a2 <- c(0,0,0,0,0,0,0)

numalpha <- length(a1)
numalpha

#creating matrix from vectors created above (proportion of variation of outcome explained by the propensity to respond)
a <- matrix(c(a1,a2), nrow=2, ncol=numalpha, byrow=TRUE)

#taking square root of proportion of variation of the propensity to respond explained by outcome. WWC refers to alpha #as similar to R², so square root is the correlation
a <- sqrt(a)

#inputting vectors of WWC values for response rates in treatment and control groups
```
p1 <- c(.9, .89, .875, .865, .85, .8, .79, .775, .765, .75, .7, .69, .675, .665, .65, .6, .59, .575, .565, .55)
p2 <- c(.9, .91, .925, .935, .95, .8, .81, .825, .835, .85, .7, .71, .725, .735, .75, .6, .61, .625, .635, .65)
numprobs <- length(p1)

#creating matrix from vectors of response rates
pp <- matrix(c(p1,p2), nrow=numprobs, ncol=2, byrow=FALSE)

#creating formula for threshold of propensity to respond
p <- 1-pp

#calling quantile function (if \( z \) is greater than the value that corresponds to a percentile of the \( z \) distribution (given \( \rho \)), then individual responds at follow-up
q <- qnorm(p)

b <- matrix(data=NA,nrow=numprobs,ncol=numalpha)
nYt <- matrix(data=NA,nrow=numprobs,ncol=numalpha)
pYt <- matrix(data=NA,nrow=numprobs,ncol=numalpha)
nYc <- matrix(data=NA,nrow=numprobs,ncol=numalpha)
pYc <- matrix(data=NA,nrow=numprobs,ncol=numalpha)
mYt <- matrix(data=NA,nrow=numprobs,ncol=numalpha)
mYc <- matrix(data=NA,nrow=numprobs,ncol=numalpha)

maxi <- as.integer(numprobs)
maxj <- as.integer(numalpha)
mini <- 1
maxi <- maxi
minj <- 1
maxj <- maxj
for (i in mini:maxi)
{
  for (j in minj:maxj)
  {
    Yt <- matrix(data=NA,nrow=n,ncol=1,byrow=FALSE)
    Yc <- matrix(data=NA,nrow=n,ncol=1,byrow=FALSE)
    #creating cumulative normal distributions for \( z \) and \( u \)
    (\( z \)=individual’s propensity to respond; \( u \)=factor unrelated to attrition [random error])
Zt <- rnorm(n)
Zc <- rnorm(n)
Ut <- rnorm(n)
Uc <- rnorm(n)

#inputting the WWC formula for outcome at follow up
for (k in 1:n)
{
  if (Zt[k] > q[i,1]) {Yt[k] <- (a[1,j] * Zt[k]) + (1-a[1,j]) * Ut[k]}
  if (Zc[k] > q[i,2]) {Yc[k] <- (a[2,j] * Zc[k]) + (1-a[2,j]) * Uc[k]}
}

nYt[i,j] <- length(which(!is.na(Yt)))
nYc[i,j] <- length(which(!is.na(Yc)))
pYt[i,j] <- nYt[i,j]/n
pYc[i,j] <- nYc[i,j]/n
mYt[i,j] <- mean(Yt, na.rm=TRUE)
mYc[i,j] <- mean(Yc, na.rm=TRUE)
#b[i,j] <- mYt[i,j]-mYc[i,j]
}

#WWC bias formula
gb <- (mYt - mYc)
gb <- round(gb,3)


Appendix F: Scenario 6 R Code Holding Control Group Relationship of The propensity to respond to Outcome at Zero and Holding Control Group Attrition Rate at 10 Percent

```r
#R version 3.0.2 (2013-09-25) -- "Frisbee Sailing"
#Copyright (C) 2013 The R Foundation for Statistical Computing
#Platform: x86_64-w64-mingw32/x64 (64-bit)
rm(list = ls())

#setting seed for random number generator so results can be replicated exactly
set.seed(18040817)

#50,000 replications
n <- 50000

#inputting vectors of WWC values for proportion of variation of outcome explained by the propensity to respond
treatment group:
a1 <- c(0.075, 0.1, 0.15, 0.2, 0.3, 0.5, 1)
#control group
a2 <- c(0,0,0,0,0,0,0)

numalpha <- length(a1)
umalpha

#creating matrix from vectors created above (proportion of variation of outcome explained by the propensity to respond)
a <- matrix(c(a1,a2), nrow=2, ncol=numalpha, byrow=TRUE)

#taking square root of proportion of variation of the propensity to respond explained by outcome. WWC refers to alpha #as similar to $R^2$, so square root is the correlation
a <- sqrt(a)
a
```
#inputting vectors of WWC values for response rates in treatment and control groups
p1 <- c(.9, .89, .875, .865, .85, .8, .79, .775, .765, .75, 
.7, .69, .675, .665, .65, .6, .59, .575, .565, .55)
p2 <- c(.9, .9, .9, .9, .9, .9, .9, .9, .9, .9, .9, .9, .9, 
.9, .9, .9, .9, .9, .9, .9)
numprobs <- length(p1)
numprobs

#creating matrix from vectors of response rates
pp <- matrix(c(p1, p2), nrow=numprobs, ncol=2, byrow=FALSE)

#creating formula for threshold of propensity to respond
p <- 1-pp

#calling quantile function (if $z$ is greater than the value that corresponds to a percentile of the $z$ distribution (given $\rho$), then individual responds at follow-up
q <- qnorm(p)

b <- matrix(data=NA, nrow=numprobs, ncol=numalpha)
nYt <- matrix(data=NA, nrow=numprobs, ncol=numalpha)
pYt <- matrix(data=NA, nrow=numprobs, ncol=numalpha)
nYc <- matrix(data=NA, nrow=numprobs, ncol=numalpha)
pYc <- matrix(data=NA, nrow=numprobs, ncol=numalpha)
mYt <- matrix(data=NA, nrow=numprobs, ncol=numalpha)
mYc <- matrix(data=NA, nrow=numprobs, ncol=numalpha)

maxi <- as.integer(numprobs)
maxj <- as.integer(numalpha)
mini <- 1
maxi <- maxi
minj <- 1
maxj <- maxj
for (i in mini:maxi)
{
  for (j in minj:maxj)
  {
    Yt <- matrix(data=NA, nrow=n, ncol=1, byrow=FALSE)
    Yc <- matrix(data=NA, nrow=n, ncol=1, byrow=FALSE)
#creating cumulative normal distributions for z and u
(z=individual’s propensity to respond; u=factor unrelated to attrition [random error])
Zt <- rnorm(n)
Zc <- rnorm(n)
Ut <- rnorm(n)
Uc <- rnorm(n)

#inputting the WWC formula for outcome at follow up
for (k in 1:n)
{
  if (Zt[k] > q[i,1]) {Yt[k] <- (a[1,j] * Zt[k]) + (1- a[1,j]) * Ut[k])}
  if (Zc[k] > q[i,2]) {Yc[k] <- (a[2,j] * Zc[k]) + (1- a[2,j]) * Uc[k])}
}

nYt[i,j] <- length(which(!is.na(Yt)))
Yt[i,j] <- nYt[i,j]/n
pYt[i,j] <- nYt[i,j]/n
mYt[i,j] <- mean(Yt, na.rm=TRUE)

mYc[i,j] <- mean(Yc, na.rm=TRUE)

#WWC bias formula
b[i,j] <- (mYt[i,j]-mYc[i,j])
}

#WWC bias formula
gb <- (mYt - mYc)
gb <- round(gb,3)
gb
Appendix G: Scenario 7 R Code Setting the Participant Group and Control Group

Relationship of The propensity to respond to Outcome as Equal

#R version 3.0.2 (2013-09-25) -- "Frisbee Sailing"
#Copyright (C) 2013 The R Foundation for Statistical Computing
#Platform: x86_64-w64-mingw32/x64 (64-bit)

rm(list = ls())

#setting seed for random number generator so results can be replicated exactly
set.seed(18040817)

#50,000 replications
n <- 50000

#inputting vectors of WWC values for proportion of variation of outcome explained by the propensity to respond
#treatment group:
a1 <- c(0.075, 0.1, 0.15, 0.2, 0.3, 0.5, 1)
#control group
a2 <- c(0.075, 0.1, 0.15, 0.2, 0.3, 0.5, 1)

numalpha <- length(a1)
numalpha

#creating matrix from vectors created above (proportion of variation of outcome explained by the propensity to respond)
a <- matrix(c(a1,a2), nrow=2, ncol=numalpha, byrow=TRUE)

#taking square root of proportion of variation of the propensity to respond explained by outcome. WWC refers to alpha #as similar to R², so square root is the correlation
a <- sqrt(a)
a
#inputting vectors of WWC values for response rates in treatment and control groups
p1 <- c(.9, .89, .875, .865, .85, .8, .79, .775, .765, .75, .7, .69, .675, .665, .65, .6, .59, .575, .565, .55)
p2 <- c(.9, .91, .925, .935, .95, .8, .81, .825, .835, .85, .7, .71, .725, .735, .75, .6, .61, .625, .635, .65)
numprobs <- length(p1)

#creating matrix from vectors of response rates
pp <- matrix(c(p1,p2), nrow=numprobs, ncol=2, byrow=FALSE)

#creating formula for threshold of propensity to respond
p <- 1-pp

#calling quantile function (if \( z \) is greater than the value that corresponds to a percentile of the \( z \) distribution (given \( \rho \)), then individual responds at follow-up
q <- qnorm(p)

b <- matrix(data=NA,nrow=numprobs,ncol=numalpha)
nYt <- matrix(data=NA,nrow=numprobs,ncol=numalpha)
pYt <- matrix(data=NA,nrow=numprobs,ncol=numalpha)
nYc <- matrix(data=NA,nrow=numprobs,ncol=numalpha)
pYc <- matrix(data=NA,nrow=numprobs,ncol=numalpha)
mYt <- matrix(data=NA,nrow=numprobs,ncol=numalpha)
mYc <- matrix(data=NA,nrow=numprobs,ncol=numalpha)

maxi <- as.integer(numprobs)
maxj <- as.integer(numalpha)
mini <- 1
maxi <- maxi
minj <- 1
maxj <- maxj
for (i in mini:maxi)
{
  for (j in minj:maxj)
  {
    Yt <- matrix(data=NA,nrow=n,ncol=1,byrow=FALSE)
    Yc <- matrix(data=NA,nrow=n,ncol=1,byrow=FALSE)
#creating cumulative normal distributions for z and u
(z=individual’s propensity to respond; u=factor unrelated
to attrition [random error])
Zt <- rnorm(n)
Zc <- rnorm(n)
Ut <- rnorm(n)
Uc <- rnorm(n)

#inputting the WWC formula for outcome at follow up
for (k in 1:n)
{
  if (Zt[k] > q[i,1]) {Yt[k] <- (a[1,j] * Zt[k]) + (1-
    a[1,j]) * Ut[k]}
  if (Zc[k] > q[i,2]) {Yc[k] <- (a[2,j] * Zc[k]) + (1-
    a[2,j]) * Uc[k]}
}

nYt[i,j] <- length(which(!is.na(Yt)))
nYc[i,j] <- length(which(!is.na(Yc)))
pYt[i,j] <- nYt[i,j]/n
pYc[i,j] <- nYc[i,j]/n
mYt[i,j] <- mean(Yt, na.rm=TRUE)
mYc[i,j] <- mean(Yc, na.rm=TRUE)
#b[i,j] <- mYt[i,j]-mYc[i,j]
}

#WWC bias formula
gb <- (mYt - mYc)
gb <- round(gb,3)
gb
Appendix H: Scenario 8 R Code Simulating Sensitivity of Model to Variation in
Reliability of Outcome Instrument Across Studies

```r
#R version 3.0.2 (2013-09-25) -- "Frisbee Sailing"
#Copyright (C) 2013 The R Foundation for Statistical Computing
#Platform: x86_64-w64-mingw32/x64 (64-bit)

rm(list = ls())

#setting seed for random number generator so results can be replicated exactly
set.seed(18040817)

#50,000 replications
n <- 50000

#inputting vectors of WWC values for proportion of variation of outcome explained by the propensity to respond
#treatment group:
a1 <- c(0.075, 0.1, 0.15, 0.2, 0.3, 0.5, 1, 1)
#control group
a2 <- c(0.05, 0.05, 0.05, 0.15, 0.2, 0.2, 1, 1)

numalpha <- length(a1)
numalpha

#creating matrix from vectors created above (proportion of variation of outcome explained by the propensity to respond)
a <- matrix(c(a1,a2), nrow=2, ncol=numalpha, byrow=TRUE)

#taking square root of proportion of variation of the propensity to respond explained by outcome. WWC refers to alpha #as similar to R^2, so square root is the correlation
a <- sqrt(a)
a

#delineating negative 1 for row 8 in control group from WWC Appendix A, Table A1
```
a[2,8] <- a[2,8]*-1
a

#inputting vectors of WWC values for response rates in
#treatment and control groups
p1 <- c(.9, .89, .875, .865, .85, .8, .79, .775, .765, .75,
   .7, .69, .675, .665, .65, .6, .59, .575, .565, .55)
p2 <- c(.9, .91, .925, .935, .95, .8, .81, .825, .835, .85,
   .7, .71, .725, .735, .75, .6, .61, .625, .635, .65)
numprobs <- length(p1)
numprobs

#creating matrix from vectors of response rates
pp <- matrix(c(p1, p2), nrow=numprobs, ncol=2, byrow=FALSE)

#creating formula for threshold of propensity to respond
p <- 1-pp

#calling quantile function (if \( z \) is greater than the value
#that corresponds to a percentile of the \( z \) distribution
#(given \( \rho \)), then individual responds at follow-up
q <- qnorm(p)

b <- matrix(data=NA, nrow=numprobs, ncol=numalpha)
nYt <- matrix(data=NA, nrow=numprobs, ncol=numalpha)
pYt <- matrix(data=NA, nrow=numprobs, ncol=numalpha)
nYc <- matrix(data=NA, nrow=numprobs, ncol=numalpha)
pYc <- matrix(data=NA, nrow=numprobs, ncol=numalpha)
mYt <- matrix(data=NA, nrow=numprobs, ncol=numalpha)
mYc <- matrix(data=NA, nrow=numprobs, ncol=numalpha)

maxi <- as.integer(numprobs)
maxj <- as.integer(numalpha)
mini <- 1
maxi <- maxi
minj <- 1
maxj <- maxj
for (i in mini:maxi)
{
  for (j in minj:maxj)
  {
    Yt <- matrix(data=NA, nrow=n, ncol=1, byrow=FALSE)
Yc <- matrix(data=NA,nrow=n,ncol=1,byrow=FALSE)

#creating cumulative normal distributions for z and u
#(z=individual’s propensity to respond; u=factor unrelated
to attrition [random error expanded for simulation of low
instrument reliability])
Zt <- rnorm(n, 0, 1.2)
Zc <- rnorm(n, 0, 1.2)
Ut <- rnorm(n, 0, 1.2)
Uc <- rnorm(n, 0, 1.2)

#inputting the WWC formula for outcome at follow up
for (k in 1:n)
{
  if (Zt[k] > q[i,1]) {Yt[k] <- (a[1,j] * Zt[k]) + (1-
a[1,j]) * Ut[k])}
  if (Zc[k] > q[i,2]) {Yc[k] <- (a[2,j] * Zc[k]) + (1-
a[2,j]) * Uc[k])}
}
nYt[i,j] <- length(which(!is.na(Yt)))
nYc[i,j] <- length(which(!is.na(Yc)))
pYt[i,j] <- nYt[i,j]/n
pYc[i,j] <- nYc[i,j]/n
mYt[i,j] <- mean(Yt, na.rm=TRUE)
mYc[i,j] <- mean(Yc, na.rm=TRUE)
#b[i,j] <- mYt[i,j]-mYc[i,j]

#WWC bias formula
gb <- (mYt - mYc)
gb <- round(gb,3)
gb