NCEE 2009-0049

U. S. DEPARTMENT OF EDUCATION

What to Do When Data Are Missing in Group Randomized Controlled Trials

ies NATIONAL CENTER FOR EDUCATION EVALUATION AND REGIONAL ASSISTANCE

Institute of Education Sciences

What to Do When Data Are Missing in Group Randomized Controlled Trials

October 2009

Michael J. Puma
Chesapeake Research Associates, LLC

Robert B. Olsen
Stephen H. Bell
Cristofer Price
Abt Associates, Inc

Abstract

This NCEE Technical Methods report examines how to address the problem of missing data in the analysis of data in Randomized Controlled Trials (RCTs) of educational interventions, with a particular focus on the common educational situation in which groups of students such as entire classrooms or schools are randomized. Missing outcome data are a problem for two reasons: (1) the loss of sample members can reduce the power to detect statistically significant differences, and (2) the introduction of non-random differences between the treatment and control groups can lead to bias in the estimate of the intervention's effect. The report reviews a selection of methods available for addressing missing data, and then examines their relative performance using extensive simulations that varied a typical educational RCT on three dimensions: (1) the amount of missing data; (2) the level at which data are missing—at the level of whole schools (the assumed unit of randomization) or for students within schools; and, (3) the underlying missing data mechanism. The performance of the different methods is assessed in terms of bias in both the estimated impact and the associated standard error.

NATIONAL CENTER FOR EDUCATION EVALUATION AND REGIONAL ASSISTANCE
Institute of Education Sciences

NCEE 2009-0049
U.S. DEPARTMENT OF EDUCATION

This report was prepared for the National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences under Contract ED-04-CO-0112/0006.

Disclaimer

The Institute of Education Sciences (IES) at the U.S. Department of Education contracted with Mathematica Policy Research to develop methods for assessing mediational analyses in education evaluations. The views expressed in this report are those of the author and they do not necessarily represent the opinions and positions of the Institute of Education Sciences or the U.S. Department of Education.

U.S. Department of Education
Arne Duncan
Secretary

Institute of Education Sciences
John Q. Easton
Director

National Center for Education Evaluation and Regional Assistance
John Q. Easton
Acting Commissioner

October 2009

This report is in the public domain. While permission to reprint this publication is not necessary, the citation should be:

Puma, Michael J., Robert B. Olsen, Stephen H. Bell, and Cristofer Price (2009). *What to Do When Data Are Missing in Group Randomized Controlled Trials* (NCEE 2009-0049). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

This report is available on the IES website at http://ncee.ed.gov.

Alternate Formats

Upon request, this report is available in alternate formats such as Braille, large print, audiotape, or computer diskette. For more information, please contact the Department's Alternate Format Center at 202-260-9895 or 202-205-8113.

Disclosure of Potential Conflicts of Interest

There are four authors for this report with whom IES contracted to develop the methods that are presented in this report. Michael J. Puma is an employee of Chesapeake Research Associates (CRA), LLC, and Robert B. Olsen, Stephen H. Bell, and Cristofer Price are employees of Abt Associates, Inc. The authors and other staff of CRA and Abt Associates do not have financial interests that could be affected by the content in this report.

Foreword

The National Center for Education Evaluation and Regional Assistance (NCEE) conducts unbiased large-scale evaluations of education programs and practices supported by federal funds; provides research-based technical assistance to educators and policymakers; and supports the synthesis and the widespread dissemination of the results of research and evaluation throughout the United States.

In support of this mission, NCEE promotes methodological advancement in the field of education evaluation through investigations involving analyses using existing data sets and explorations of applications of new technical methods, including cost-effectiveness of alternative evaluation strategies. The results of these methodological investigations are published as commissioned, peer reviewed papers, under the series title, Technical Methods Reports, posted on the NCEE website at http://ies.ed.gov/ncee/pubs/. These reports are specifically designed for use by researchers, methodologists, and evaluation specialists. The reports address current methodological questions and offer guidance to resolving or advancing the application of high-quality evaluation methods in varying educational contexts.

This NCEE Technical Methods report examines how to address the problem of missing data in the analysis of data in Randomized Controlled Trials (RCTs) of educational interventions, with a particular focus on the common educational situation in which groups of students such as entire classrooms or schools are randomized. Missing outcome data are a problem for two reasons: (1) the loss of sample members can reduce the power to detect statistically significant differences, and (2) the introduction of non-random differences between the treatment and control groups can lead to bias in the estimate of the intervention’s effect. The report reviews a selection of methods available for addressing missing data, and then examines their relative performance using extensive simulations that varied a typical educational RCT on three dimensions: (1) the amount of missing data; (2) the level at which data are missing─at the level of whole schools (the assumed unit of randomization) or for students within schools; and, (3) the underlying missing data mechanism. The performance of the different methods is assessed in terms of bias in both the estimated impact and the associated standard error.

Table of Contents

1. Overview and Guidance

A. Introduction

Most statistics textbooks provide lengthy discussions of the theory of probability, descriptive statistics, hypothesis testing, and a range of simple to more complex statistical methods. To illustrate these discussions, the authors often present examples with real or fictional data─tidy tables of observations and variables with values for each cell. Although this may be entirely appropriate to illustrate statistical methods, anyone who does "real world" research knows that data are rarely, if ever, so complete. Some study participants may be unavailable for data collection, refuse to provide data, or be asked a question which is not applicable to their circumstances. Whatever the mechanism that causes the data to be missing, it is a common problem in almost all research studies.

This report is designed to provide practical guidance on how to address the problem of missing data *in the analysis of data in Randomized Controlled Trials (RCTs) of educational interventions,* with a particular focus on the common educational situation in which groups of students such as entire classrooms or schools are randomized (called Group Randomized Trials, GRTs). The need for such guidance is of growing importance as the number of educational RCTs has increased in recent years. For example, the ten Regional Educational Laboratories (RELs) sponsored by the Institute of Education Sciences (IES) of the U.S. Department of Education are currently conducting 25 RCTs to measure the effectiveness of different educational interventions,[1] and IES has sponsored 23 impact evaluations that randomized students, schools, or teachers since it was established in 2002.[2]

This report is divided into four chapters. Following a brief overview of the missing data problem, this first chapter provides our overall guidance for educational researchers based on the results of extensive data simulations that were done to assess the relative performance of selected missing data strategies within the context of the types of RCTs that have been conducted in education. Chapter 2 sets the stage for a discussion of specific missing data strategies by providing a brief overview of the design of RCTs in education, the types of data used in impact analysis, how these data can be missing, and the analytical implications of missing data. Chapter 3 describes a selection of methods available for addressing missing data, and Chapter 4 describes the simulation methodology and the statistical results that support the recommendations presented in this chapter. Appendices provide additional details on the simulation methods and the statistical results.

[1] The Regional Educational Laboratories serve the states in their regions with research and technical assistance, including both original studies—of which the RCTs are the largest—and syntheses of existing research. For more information on the RELs, see http://ies.ed.gov/ncee/edlabs/.

[2] See ongoing and completed evaluation studies sponsored by IES's National Center for Education Evaluation and Regional Assistance at http://ies.ed.gov/ncee/projects/evaluation/index.aspyear.

B. Missing Data and Randomized Trials

The purpose of a randomized controlled trial (RCT) is to allow researchers to draw causal conclusions about the effect, or "impact," of a particular policy-relevant intervention (U.S. Department of Education, 2009). For example, if we wanted to know how students do when they are taught with a particular reading or math curriculum, we could obtain test scores before and after they are exposed to the new mode of instruction to see how much they learned. But, to determine if this intervention **caused** the observed student outcomes we need to know how these **same** students would have done had they **not** received the treatment.[3] Of course, we cannot observe the same individuals in two places at the same time. Consequently, the RCT creates equivalent groups by **randomly assigning** eligible study participants either to a ***treatment group***, that receives the intervention under consideration, or to a ***control group***, that does not receive the particular treatment but often continues with "business as usual," e.g., the mode of instruction that would be used in the absence of a new math curriculum.[4] [5] Because of the hierarchical way in which schools are organized, most education RCTs randomize groups of students─entire schools or classrooms─rather than individual children to study conditions. In these GRTs the treatment is typically delivered at the group or cluster level but the primary research interest is the impact of the selected treatment on student outcomes, although it is also not uncommon to look for intermediate impacts on teachers or schools.

The advantage of the RCT is that if random assignment is properly implemented with a sufficient sample size, treatment group members will not differ in any systematic or unmeasured way from control group members except through their access to the intervention being studied (the groups are equivalent both on observable and unobservable characteristics). It is this elimination of confounding factors that allows us to make unbiased causal statements about the effect of a particular educational program or intervention by contrasting outcomes between the two groups.

However, for an RCT to produce unbiased impact estimates, the treatment and control groups must be equivalent in their composition (in expectation) not just at the point of randomization (referred to as the "baseline" or pretest point), but also at the point where follow-up or outcome data are collected. Missing outcome data are a problem for two reasons: (1) the loss of sample members can reduce the power to detect statistically significant differences, and (2) the introduction of non-random differences between the treatment and control groups can lead to bias in the estimate of the intervention's effect. The seriousness of the potential bias is related to the overall magnitude of the missing

[3] For simplicity we use an example of a simple two-group design with a single treatment and control group. Real world RCTs can include a variety of combinations including multiple treatment arms, and designs in which there is no control group, i.e., the study compares outcomes across different treatments.

[4] In some RCTs, the counterfactual represented by the control group reflects the conditions that would prevail in the absence of the intervention being tested. This counterfactual condition is often, and perhaps misleadingly, referred to as "business as usual." In other RCTs, the study is designed to compare the impacts of two alternative interventions.

[5] In most cases, study participants are randomly assigned on an equal basis to the treatment and control groups. However, there are situations in which there is good reason to use unequal allocation, for example, where there is strong resistance to placing participants into the control group. Such variation from a 50:50 allocation will result in some loss of statistical power, but this is generally modest in magnitude.

data rate, and the extent to which the likelihood of missing data differs between the treatment and control groups. For example, according to the ***What Works Clearinghouse***[6] the bias associated with an overall attrition rate of ten percent and a differential treatment-control group difference in attrition rates of five percent can be equal to the bias associated with an overall attrition rate of 30 percent and a differential attrition rate of just two percent.

Therefore, in a perfect world, the impact analysis conducted for an RCT in education would include outcomes for all eligible study participants defined at the time of randomization. However, this ideal is rarely ever attained. For example, individual student test scores can be completely missing because of absenteeism, school transfer, or parental refusal for testing. In addition, a particular piece of information can be missing because respondents refuse to answer a certain test item or survey question, are unable to provide the requested information, inadvertently skip a question or test item, or provide an unintelligible answer. In an education RCT, we have to also concern ourselves with missing data at the level of entire schools or classrooms if randomly assigned schools or classrooms opt either out of the study completely or do not allow the collection of any outcome data.

As demonstrated by Rubin (1976, 1987), the process through which missing data arise can have important analytical implications. In its most innocuous form−a category that Rubin calls ***Missing Completely at Random (MCAR)***−the mechanism that generates missing data is a truly random process unrelated to any measured or unmeasured characteristic of the study participants. A second category−***Missing at Random (MAR)***−is one in which missingness is random conditional on the observed characteristics of the study sample. For example, the missing data would be MAR if missingness on the post-test score were related to gender, but conditional on gender—that is, among boys or among girls—the probability of missing data is the same for all students. Typically, if one can reasonably assume that missing data arise under either the conditions of MCAR or MAR the missing data problem can be considered "ignorable," i.e., the factors that cause missingness are unrelated, or weakly related, to the estimated intervention effect. In some situations, however, one cannot reasonably assume such ignorability–a category that Rubin calls ***Not Missing at Random (NMAR.*** [7]

Within the context of an RCT, if the missing data mechanism differs between the treatment and control groups, dropping cases with missing data may lead to systematic differences between the experimental groups which can lead to biased impact estimates. Furthermore, even if the missing data mechanism is the same for the treatment and control groups, we may still be concerned if certain types of teachers or students are under- or over-represented in the analysis sample. For example, if the impacts for underrepresented groups are higher than the average impact, then the impact estimates will be biased downward; if the impacts for underrepresented groups are lower than the average impact, the impact estimates will be biased upward.

[6] U.S. Department of Education, (2008).

[7] The category is also described in the literature as non-ignorable non-response (NINR).

C. Missing Data Methods

As noted above, missing data is a common problem in educational evaluations. For example, in impact evaluations funded by the National Center for Educational Evaluation and Regional Assistance (NCEE), student achievement outcomes are often missing for 10-20 percent of the students in the sample (Bernstein, et al., 2009; Campuzano, et al., 2009; Constantine, et al., 2009; Corrin, et al., 2008; Gamse, et al., 2009; Garet, et al., 2008; and Wolf, et al., 2009). Strategies used to address missing data in education RCTs range from simple methods like listwise deletion (e.g., Corrin, et al., 2009), to more sophisticated approaches like multiple imputation (e.g., Campuzano, et al., 2009). In addition, some studies use different approaches to addressing missing covariates and missing outcomes, such as imputing missing covariates but re-weighting complete cases to address missing outcomes (e.g., Wolf, et al., 2009).

Despite the prevalence of the missing data challenge, there is no consensus on which methods should be used and the circumstances under which they should be employed. This lack of common standards is not unique to education research, and even areas where experimental research has a long history, like medicine, are still struggling with this issue. For example, guidance from the Food and Drug Administration (FDA) on the issue of missing data in clinical trials indicates that “A variety of statistical strategies have been proposed in the literature…(but) no single method is generally accepted as preferred” (FDA, 2006, p.29).

The selection of the methods that are the focus of this report was based on a review of several recent articles by statistical experts seeking to provide practical guidance to applied researchers (Graham, 2009; Schafer & Graham, 2002; Allison, 2002; and Peugh & Enders, 2004). Specifically, this report examines the following analysis strategies for dealing with the two types of missing data that are of primary importance when analyzing data from an educational RCT─missing outcome or post-test data and missing baseline or pretest data:

- **Appropriate for Missing Pretest Data Only:**
  - **Dummy Variable Adjustment─**setting missing cases to a constant and adding “missing data flags” to the impact analysis model.
- **Appropriate for Missing Post-test Data Only:**
  - **Weighting─**re-balancing the analysis sample to account for the loss of study participants.
  - **Fully-Specified Regression Models─**adding to the impact analysis model terms that interact the covariates with the treatment indicator.[8]

[8] This method may have only been used in the single study for which it was developed (Bell & Orr, 1994). However, we decided to include it in our review because it reflects a fundamentally different approach to missing data focusing on the re-specification of the analysis model.

- **Appropriate for Both Types of Missing Data:**
  - **Imputation Methods**─filling in missing values using one of four methods, single mean imputation, single non-stochastic regression imputation, single stochastic regression imputation, and multiple stochastic regression imputation.
  - **Maximum Likelihood─EM Algorithm with Multiple Imputation**─a statistical estimation method that tries to find the population parameters that are most likely to have produced a particular data sample, using all of the available observations including those with missing data.
  - **Selection Modeling and Pattern Mixture Modeling**─two attempts to deal with the NMAR situation by statistically modeling the missing data mechanism.

In the discussions that follow, we intentionally include methods that are commonly criticized in the literature─listwise deletion and simple mean value imputation─for two reasons. First, the use of these methods is widespread in education. For example, a review of 545 published education studies by Peugh & Enders (2004) showed a nearly exclusive reliance on deletion as a way to deal with missing data. Second, because we are focusing on RCTs, we wanted to understand how different missing data strategies performed within this unique context, including commonly used but criticized methods.

D. Guidance to Researchers

Basis for the Recommendations

Our recommendations for dealing with missing data in Group Randomized Trials in education are based on the results of extensive statistical simulations of a typical educational RCT in which schools are randomized to treatment conditions. As discussed in Chapter 4, selected missing data methods were examined under conditions that varied on three dimensions: (1) the amount of missing data, relatively low (5% missing) vs. relatively high (40% missing); (2) the level at which data are missing─at the level of whole schools (the assumed unit of randomization) or for students within schools; and, (3) the underlying missing data mechanisms discussed above (i.e., MCAR, MAR, and NMAR).

The performance of the selected missing data methods was assessed on the basis of the bias that was found in both the estimated impact and the associated estimated standard, using a set of standards that were developed from guidance currently in use by the U.S. Department of Education's *What Works Clearinghouse* (see Chapter 4 and Appendix E).

The recommendations that are provided below are based on the following criteria:

- In general, we recommend avoiding methods for which the simulations indicated a bias, in either the magnitude of the estimated impact or its associated standard error, that exceeded 0.05 standard deviations of the outcome measure (a standard developed on the basis of the current WWC guidance on sample attrition).

- The recommendations are based on the simulation results in which data were missing for 40 percent of students or schools. This is because the alternative set of simulation results in which data were missing for five percent of either students or schools showed that all of the tested methods produced results that fell within the WWC-based standard. We recognize that many studies in education will have lower missing data rates than 40 percent. However, our recommendations are designed to be conservative─avoiding methods that produce a large amount of bias when the missing data rate is 40 percent will reduce the likelihood that that an evaluation will suffer from bias of this magnitude if the missing data rate is less than 40 percent.
- We provide recommendations separately for missing pretest scores[9] and for missing post-test scores (covariates):
  - **For missing pretests**, recommended methods must have produced bias below the established thresholds for both impacts and standard errors. Bias in either estimate can lead to biased t-statistics and invalid statistical inference. Therefore, we only recommend methods that produce estimates with low bias for both impacts and standard errors.

    In addition, we only recommend methods that produced estimates with low bias *in all three scenarios*, i.e., MCAR, MAR, and NMAR. Because each scenario reflects a different missing data mechanism, and the missing data mechanism is never known in actual studies, methods that produced estimates with low bias in all three scenarios can be considered "safer choices" than methods that produced estimates with low bias in some but not all of the scenarios.
  - **For missing post-test scores**, none of the methods produced impact estimates with bias of less than 0.05 when missing data are NMAR. Therefore, requiring methods to produce estimates with low bias in all three scenarios would have left us with no methods to recommend to analysts facing missing outcome data in their studies. As a consequence, for missing post-test scores, we recommend methods that produced estimates meeting our performance standard when missing data are both MCAR and MAR, recognizing that even recommended methods may produce higher levels of bias under the NMAR condition.

[9] Although the detailed results provided in Chapter 4 and Appendix D include models that include or exclude the pretest covariate, our recommendations are based on the simulations in which pretest scores were available. When pretest scores were not available, some of the methods we recommend produced impact estimates with bias of greater than 0.05 standard deviations. Our simulation results suggest that to produce impact estimates with the bias below this threshold depends on both the choice of methods and the availability of data on important covariates.

Recommendations

Missing Pretest Scores Or Other Covariates

When pretest scores or other covariates are **missing for students within schools** in studies that randomize schools the simulation results lead us to recommend the use of the following missing data methods:

- Dummy variable adjustment,
- Single stochastic regression imputation,
- Multiple stochastic regression imputation (i.e., "multiple imputation"), and
- Maximum Likelihood─EM algorithm with multiple imputation.

In this context, we would **not** recommend the use of three methods that produced impact estimates with bias that exceeded 0.05 in one or more of our simulations: case deletion, mean value imputation, and single non-stochastic regression imputation.

Alternatively, when data on baseline variables are **missing for entire schools**, the simulation results lead us to recommend the use of the following methods:[10]

- Case deletion,
- Dummy variable adjustment,
- Mean value imputation,
- Multiple stochastic regression imputation, and
- Maximum Likelihood─EM algorithm with multiple imputation.

We would **not** recommend the use of two methods that produced standard error estimates with bias in at least one of our simulations that exceeded the WWC-based threshold: single non-stochastic regression imputation and single stochastic regression imputation.

Across the two scenarios─i.e., situations when pretest or covariate data are missing either for students within schools or for entire schools─three methods were found to be consistently recommended so are likely to be the best choices:

- Dummy variable adjustment,
- Multiple stochastic regression imputation, and
- Maximum Likelihood─EM algorithm with multiple imputation.

It is important to stress that these recommendations are specific to situations in which the intent is to make inferences about the coefficient on the treatment indicator in a group randomized trial. If, for example, an analyst wanted to make inference about the relationship between pretest and post-test scores, and there were missing values on the pretest scores, we would not recommend use of the dummy variable approach. With this method, the estimate of the coefficient on the pretest score is likely to be biased, as has been described in previous literature. But when the interest is on the estimate of the treatment effect, the dummy variable approach yields bias in the coefficient of interest— the estimated treatment effect— that falls within the acceptable range as we defined it for

[10] Note that the "simple" weighting approach cannot be applied when data are missing for entire schools, because it involves weighting up respondents from a given school to represent nonrespondents from that school; with school-level missing data there are no respondents to use for this purpose.

these simulations, and is similar in magnitude to the biases obtained from the more sophisticated methods.

Missing Post-Test Scores Or Other Outcome Variables

When data on outcome variables are **missing for students within schools** in studies that randomize schools, the simulation results lead us to recommend the use of the following methods:

- Case deletion,
- Single non-stochastic regression imputation,
- Single stochastic regression imputation,
- Multiple stochastic regression imputation,
- Maximum Likelihood─EM algorithm with multiple imputation,
- "Simple" weighting approach using the inverse of observed response rates,
- "Sophisticated" weighting approach that involved modeling non-response to create weights, and
- Fully-specified regression models with treatment-covariate interactions.

We would **not** recommend using mean value imputation because it was the only method that produced impact estimates with bias that exceeded 0.05 under the MAR scenario.

When data on dependent variables are **missing for entire schools**, the simulation results lead us to recommend the use of the following methods:

- Case deletion,
- Multiple stochastic regression imputation,
- Maximum Likelihood─EM algorithm with multiple imputation,
- Sophisticated weighting approach, and
- Fully specified regression model with treatment-covariate interactions.

We would **not** recommend the use of the three methods that produced standard error estimates with bias that exceeded the WWC-based threshold: mean value imputation, single non-stochastic regression imputation, and single stochastic regression imputation.

Across the two scenarios─i.e., situations when data on post-test or outcome data are missing either for students within schools or for entire schools─five methods were found to be consistently recommended so are likely to be the best choices:

- Case deletion,
- Multiple stochastic regression imputation,
- Maximum Likelihood─EM algorithm with multiple imputation,
- Sophisticated weighting approach, and
- Fully-specified regression models.

In addition, we recommend that if post-test scores are missing for a high fraction of students or schools (e.g., 40%), analysts should control for pretest scores in the impact model if possible. In our simulations, controlling for pretest scores by using them as regression covariates reduced the bias in the impact estimate by approximately 50 percent, and this finding was robust across different scenarios and different methods.[11]

[11] For example, when missing post-tests depended on the values of the post-test scores, and data were missing for 40 percent of students, including the pretest score as a covariate in the models reduced the bias from 0.124 to 0.67 for case

As a final note, the recommendations provided above indicate that some methods that are easy to implement performed similarly to more sophisticated methods. In particular, where pretest scores were missing for either students or entire schools, the dummy variable approach performed similarly to the more sophisticated approaches and was among our recommended methods. And when post-test scores were missing for either students or entire schools, case deletion was among our recommended approaches. Consequently, we suggest that analysts take the ease with which missing data can be handled, and the transparency of the methods to a policy audience, into consideration when making a final choice of an appropriate method.

Other Suggestions For Researchers

In addition to the recommendations that we derive from the simulation results presented in this report, we also recommend that all researchers adhere to the following general analysis procedures when conducting any education RCT:

- **What to do during the analysis planning stage?** Researchers should carefully describe, and commit to, a plan for dealing with missing data **before** looking at preliminary impact estimates or any outcome data files that include the treatment indicator variable. Committing to a design, and then sticking to it, is fundamental to scientific research in any substantive area. This is best accomplished by publishing the missing data plan prior to collecting the outcome data. If the original plan then fails to anticipate any of the missing data problems observed in the data, the plan should be updated and revised before any impact estimates are produced.

  Researchers should also consider conducting sensitivity analysis to allow readers to assess how different the estimated impacts might be under different assumptions or decisions about how to handle missing data (see Chapter 3 for a discussion of one approach involving placing "best and worst case" bounds around the estimated impacts). These planned sensitivity tests should be specified ahead of time in the analysis plan.

- **What information should the impact report provide about missing data?** In their impact report, researchers should report missing data rates by variable, explain the reasons for missing data (to the extent known), and provide a detailed description of how missing data were handled in the analysis, consistent with the original plan.[12] Impact reports should also provide key descriptive statistics for the study sample, including: (1) differences between the treatment and control group on baseline characteristics both at the point of random assignment and for the impact analysis sample (excluding, of course, any missing data imputations); and, (2) differences in baseline characteristics between treatment group respondents and non-respondents (i.e., those with and without outcome data), and similarly between control group respondents and non-respondents.

deletion, and it reduced the bias from 0.122 to 0.061 for multiple stochastic regression imputation (see Appendix D, Table III.b.1).

[12] Additional guidelines for reporting missing data patterns and procedures appear in Burton & Altman (2004).

Final Caveats

Readers are cautioned to keep in mind that these simulation results are specific to a particular type of evaluation—an RCT in which *schools* are randomized to experimental conditions. Whether the key findings from these simulations would apply in RCTs that randomize students instead of schools is an open question that we have not addressed in this report. In addition, it is not clear whether the findings from our simulations would be sensitive to changes in the key parameter values that we set in specifying the data generating process and the missing data mechanisms. Finally, we could not test all possible methods to address missing data. These limitations notwithstanding, we believe these simulations yield important results that can help inform decisions that researchers need to make when they face missing data in conducting education RCTs.

Finally, despite all of the insights gained from the simulations, we cannot propose a fully specified and empirically justified decision rule about which methods to use and when. Too often, the best method depends on the purpose and design of the particular study and the underlying missing data mechanism that cannot, to the best of our knowledge, be uncovered from the data.

2. Randomized Controlled Trials (RCTs) in Education and the Problem of Missing Data

A. Why Conduct RCTs?

The Randomized Controlled Trial (RCT) has long been a mainstay of medical research to examine the effectiveness of different types of health care services (e.g., approaches to medical and nursing practice) as well as technologies such as pharmaceuticals and medical devices. In recent years, RCTs have become the "gold standard" for social policy evaluation in a wide range of areas including education (U.S. Department of Education, 2008).

RCTs are well designed to solve the classic problem of causal inference (commonly referred to as the "Rubin Causal Model") that arises when we can observe outcomes for individuals in the group that receive the treatment but we cannot observe what would have happened if these ***same*** individuals had not received the selected intervention (e.g., Imbens & Wooldridge, 2009). For example, we cannot observe how the same class of students would have performed on a standardized test if they had been taught using a different curriculum or teaching method. All we can observe for the children is how they did when taught by their current teacher with whatever that entails in terms of the curriculum or pedagogical approach. To address this problem, random assignment produces a control group that differs systematically from the treatment group in only one way—receipt of the intervention being evaluated. Therefore, the control group yields information on how the treatment group would have fared under the counterfactual, or "untreated," condition.

As discussed in Chapter 1, the advantage of the RCT is that if random assignment is properly implemented (i.e., the process is truly random) with a sufficient sample size, program participants are not expected to differ in any systematic or unmeasured way from non-participants except through their access to the new instructional program.[13] By eliminating the effect of any confounding factors, randomization allows us to make causal statements about the effect of a particular educational program or intervention, i.e., observed outcome differences are caused by exposure to the treatment. In fact, with a randomized design, if one has complete outcome data a simple comparison of treatment-control group average outcomes yields an unbiased estimate of the impact of the particular program or intervention on the study participants.

This certainty of attribution to the right causal factor can never be achieved if schools and staff make their own choices regarding, for example, the type of instruction used for mathematics. Too many things about the schools, teachers, and students could potentially differ, and this can undermine our ability to reliably attribute observed outcome differences to the single causal factor─the treatment condition. Although researchers have suggested a large number of non-experimental methods for achieving the same purpose such as multivariate regression, selection correction methods (Heckman & Hotz,

[13] More precisely, there may be differences between individuals in the two groups, but other than the influence of the intervention the expected value of these differences is zero. That is, the bias that may result from individual selection is removed by random assignment.

1989), and propensity score methods (Rosenbaum & Rubin, 1983), a long line of literature, including recent analyses by Bloom, et al. (2002), Agodini & Dynarski (2001), and Wilde & Hollister (2002), suggests that none of these methods provides causal attribution matching the reliability of random assignment.

B. RCTs in Education

RCTs have been used in education to estimate the impacts of a wide range of interventions, including evaluations of broad federal programs such as Upward Bound (Seftor, et al., 2009) and Head Start (Puma, et al., 2005), school reform initiatives such as Success for All (Borman, et al., 2007) and Comer's School Development Program (Cook, et al., 1999), and subject-specific instructional programs such as Accelerated Reader (Ross, et al., 2004; Bullock, 2005) and Connected Mathematics (REL-MA, 2008). In the case of instructional programs, the direct treatment that is being manipulated often involves training teachers in a new curriculum or instructional practice, and the trained teachers are then expected to implement the new approach in their classrooms. Because the primary interest of such RCTs is the impact on student learning, the actual treatment includes both the training itself plus how teachers, in fact, implement the new instructional method, including any real world adaptations and distortions of the expected intervention. For example, among the 25 RCTs currently being conducted by the IES-funded Regional Educational Labs (RELs), 20 are testing different models of instructional practice that include a teacher professional development component.[14]

Outside the field of education, it is common to randomly assign individuals to the treatment and control groups (e.g., individual patients who do or do not get a new drug regimen), but these are less common in the field of education. More typically, researchers conduct Group Randomized Trials (GRTs) in which the units of random assignment are intact groups of students─either entire schools or individual teachers and their classrooms─but the primary interest of the study is typically the impact of the selected treatment on student-level outcomes (although it is not uncommon to look for intermediate impacts on schools or teachers). This leads to a hierarchical, or nested, research design in which the units of observation are members of the groups that are the actual units of random assignment. As Murray (1998) describes, the groups that are the units of random assignment are not generally formed at random but there is some connection that creates a correlation among the individuals within the group. For example, 3rd grade students in the class of a particular teacher are likely to share a variety of similar characteristics. (This correlation has important analytic implications for the simulations described in Chapter 4.)

GRT designs are well illustrated by the collection of 25 experimental studies currently being conducted by the RELs of which 17 randomly assigned entire schools and six assigned classes/teachers within schools (two assigned individual students). The studies typically involve a single cohort of study participants, but five have multiple annual cohorts. Most (22) are conducting follow-up testing of students using some form of standardized test to measure student achievement, while eight are collecting scores on state assessments from administrative records (either as the sole outcome measure or in

[14] For a review of the REL studies currently underway see http://ies.ed.gov/ncee/edlabs/relwork/index.asp.

conjunction with study-administered testing). Some (5) are also measuring student non-achievement outcomes (e.g., course-taking, instructional engagement) from student surveys or school administrative records. Because many of the interventions involve teacher professional development, several (9) include measures of teacher practice and two have included teacher tests to gauge teacher knowledge. Follow-up data collection is typically conducted at a single point in time, approximately one year after randomization, but can also include multiple outcome testing points; essentially all of the studies collected baseline data to improve the precision of the impact estimates and identify student subgroups of particular interest (e.g., pre-intervention test scores, student demographic characteristics).

C. Defining the Analysis Sample

Because the unit of assignment is usually the school, classroom, or teacher, while the unit of analysis is the student, multi-level modeling is the typical approach to impact estimation in group RCTs to account for the associated clustering.[15] These models include a dummy variable to distinguish the treatment group from the control group at the appropriate level, depending on the unit of assignment,[16] and control variables at the student level, such as pre-intervention test scores, to increase the precision of the impact estimates. The estimated coefficient on the treatment indicator provides the study's estimate of the average effect of the intervention on all students randomized. Referred to as the "intent-to-treat" (ITT) effect, this estimate would, for example, capture the impact of a policy which *made professional development available* to teachers regardless of whether all of the individual teachers actually took part in the training. In other words, the ITT effect captures the impact of the **offer** of a particular intervention, not the impact of a school or a teacher actually participating in the intervention.

There are at least two reasons to focus on estimating the ITT effect. First, it is the true experimental impact estimate because all treatment and control group members are included in the analysis. Second, the ITT effect often reflects the impact of the feasible policy option─making a particular program available to a specified set of intended participants. That is, a program can be made available but whether it is implemented as intended, or whether all of the targeted participants actually get the intervention, is difficult if not impossible to control. For example, consider an RCT on teacher professional development that could inform a state policy decision on whether to offer a particular type of training to some or all schools in the state. In this case, state policy makers would benefit from evidence on the impacts of offering the professional

[15] Studies that fail to account for this nesting structure will typically underestimate the standard errors of the impact estimates, making them more likely to find significant impact estimates when the true impacts are zero. For more information, see the What Works Clearinghouse's on-line tutorial concerning the mismatch between the unit of assignment and the unit of analysis (http://ies.ed.gov/ncee/wwc/pdf/mismatch.pdf).

[16] For example, RCTs in which the school is the unit of assignment should include a treatment indicator to identify treatment schools in the school-level equation of the model.

development to schools—not on the effects of schools accepting the offer, teachers receiving the training, or other factors that are beyond the control of state policymakers.[17]

If all RCTs in education need to do a good job of estimating the ITT effect, what sample becomes the target for data collection? *In estimating the ITT effect, we need to collect data on and analyze all students and schools that were randomly assigned.* For example, it would be convenient to simply exclude students from the study if they move to a school that cannot provide data, but there *may* be systematic differences between the "stayers" in the two groups, especially if the treatment affects the probability of remaining in the school. Therefore, excluding the "movers" from the sample—or removing any other group from the sample on the basis of a factor that could have been affected by the treatment—undermines the internal validity that randomization was designed to ensure.[18]

Even treatment group members who do not get the intervention have to be part of the impact analysis sample. ITT analysis does not allow them to be omitted from the analytical sample, because their counterparts in the control group cannot be identified and similarly excluded to maintain the equivalence of the two groups.

Therefore, it is important to either obtain data on the full randomized sample, or, when this is not feasible, to select appropriate methods for addressing missing data. Hence, regardless of the research goal, the missing data methods in this report should be applied to the full randomized sample, and in the chapters that follow the different methods are assessed in terms of their ability to successfully deal with the potential bias that may be introduced when data are missing.

D. How Data Can Become Missing

Most statistical textbooks rarely deal with the real world situation in which study participants are entirely absent from the data set, or situations where particular study participants lack data on one or more analytical variables. For example, as defined in the Stata manual, "Data form a rectangular table of numeric and string values in which each row is an observation on all of the variables and each column contains the observations on a single variable."[19] This "default" solution to missing data−referred to as *listwise* or *casewise deletion,* or *complete case analysis*−is a common feature of most statistical packages. This approach is simple to implement, but it can reduce the size of the available sample and associated statistical power, and, as discussed later in this report, may introduce bias in the impact estimate.

[17] The importance of the ITT estimate notwithstanding, there can also be interest in the impact of the intervention on the teachers or students *who are directly affected by the treatment*, referred to as the effect of the Treatment on the Treated (TOT). For example, in many education RCTs we may be interested primarily in the impact of actually receiving a particular type of professional development training on teacher performance, as measured by student test scores. However, even in this case, it is generally accepted that obtaining an unbiased estimate of the TOT effect begins by reliably estimating the ITT effect (see the recent discussion in Gennetian, et al., 2005).

[18] This is true even when statistical tests show no statistically significant *impact* of the intervention on mobility (i.e., no treatment minus control group difference in the proportion who move and thus lack follow-up data). The proportions may differ but the test lack sufficient power to *prove* that they do, or the proportions may match even though the *composition* of the movers—and hence the composition of the non-movers actually analyzed—is affected by the intervention and differs between the treatment and control group analysis samples.

[19] Stata Release 9, User's Guide, Stata Press, College Station, Texas (p 117).

The most obvious way that data can become missing is a failure to obtain any information for a particular participant or observation, a situation called "unit non-response" in the survey literature. In education this can occur for several reasons. For example, individual student test scores may be missing because parents did not consent to have their child tested, students were absent or had transferred to another school, the child was exempted from taking the test (e.g., because of a disability or limited English language ability), or the student's classroom was unavailable for testing (e.g., a fire drill took place). In the case of test scores from administrative records, the school or district may have been unwilling or unable to provide data. Teacher data may be missing because of a refusal to complete a survey or take a test, extended absence from school, or transfer to another school.

In addition to the complete absence of data for a particular randomized study participant, an often more common problem is "item non-response" where respondents refuse to answer a particular question, are unable to provide the information ("don't know"), inadvertently skip a question or test item, or provide an unintelligible answer. Sometimes data may be "missing" by design in survey data because a particular question is not applicable.[20] Or certain questions may be skipped to reduce burden on individual respondents, who receive only a subset of the full set of possible questions (called "matrix sampling").

Longitudinal studies in which data are collected from study participants at multiple time points, or "waves," present different missing data possibilities. For example, consider a study in which data were collected at four separate time points and data are available as shown in the example below:

Student	***Wave 1***	***Wave 2***	***Wave 3***	***Wave 4***
A	X	X	X	X
B	X		X	X
C	X	X		X
D		X	X	X
E	X			X
F	X	X		

In this example, Student A was tested at all waves. Students B-E provided incomplete data, i.e. data were not obtained some data points (wave non-response). Student F only provided data at the first and second time points – this child was a "drop out" from the study (often called study attrition) because, for example, he left the study school. As will be discussed in the next chapter, these patterns of wave non-response can provide some opportunities for imputing or otherwise adjusting for the missing time points in the

[20] Take, for example, the question, "Do you own or rent?" If a respondent answers "Rent" to this screener question, he is then asked the amount of the monthly rent payment; if the respondent is not a renter, then the item is skipped. Non-responses to such skipped items are not generally considered to be missing data.

sequence, to the extent that outcome measures for a given student are likely to be correlated over time.[21]

E. The Missing Data Problem

As discussed in Chapter 1, there are two potential problems that can result from missing data.[22] First, if the missing data mechanism is different for the treatment and control group, dropping cases with missing data can introduce systematic differences which can, in turn, lead to biased impact estimates. Second, even if the missing data mechanism is the same for the treatment and control groups, we may still be concerned about missing data if certain types of teachers or students are more likely to have missing data and thus are under-represented in the analysis sample. If the impact of the educational intervention varies, this can lead to biased impact estimates: for example, if the impacts for underrepresented groups are higher than the average impact, then the average impact estimates will be biased downward; alternatively, if the impacts for underrepresented groups are lower the average impact estimates will be biased upward.

We believe that both of these problems are serious. In the first case, the seriousness of the problem is probably noncontroversial: if missing data produces systematic differences between the complete cases in the treatment group and the complete cases in the control group, then the impact estimates will be biased. However, the second problem may warrant additional consideration. In many RCTs in education, the schools in the sample constitute a sample of convenience: they are not selected randomly and thus cannot formally be considered representative of any larger population (e.g., Bernstein, et al., 2009; Constantine, et al., 2009; and Garet, et al., 2008). Therefore, some analysts may argue that if the study is not designed to produce externally valid estimates, we should be less concerned about missing data problems that make the analysis sample "less representative." However, some RCTs in education do in fact select schools or sites randomly. Furthermore, in those RCTs that select a nonrandom sample of convenience—schools willing to participate in the study—the study's goal is presumably to obtain *internally valid estimates of the intervention's impact for that sample of schools*. If missing data problems lead to a sample of students with complete data in those schools that is not representative of all students in those schools, we believe this is a problem that should be addressed.

The bias that can be introduced by missing data can most easily be understood by considering a typical education RCT. For example, consider a study for which one has a primary student-level outcome variable, Y, such as a student assessment in reading, a treatment indicator *"Trt"* (*Trt*=1 if assigned to the treatment group, and =0 if assigned to the control group, which we assume is always known), and a set of covariates, X_1 through X_n, all measured at the time of, or prior to, random assignment (e.g., a student's prior assessment score and other demographic variables). Although we could have missing

[21] A technique that is often used in medical trials–"last observation carried forward (LOCF)"—uses the last observed data point as the final time point measure for patients who drop out of the study. We do not include this method in this report because it is highly criticized in the medical literature and other methods we do examine are superior. For more information see Wood, White & Thompson, 2004.

[22] Appendix A examines the problem of missing data from the additional perspective of omitted variable bias.

data for the outcome variable *Y* or for any of the control variables X_1 through X_n, for simplicity let us consider cases where only the outcome variable is missing for some observations.

As discussed in Chapter 1, the most innocuous form of missing data in the Rubin framework is called ***Missing Completely at Random (MCAR)***. In our example, this situation would hold if the probability of the outcome test score being missing is unrelated to the student's "true" test score (i.e., students who would score higher or lower at the point of outcome testing are not more or less likely to be missing) or to any of the other important measured student characteristics (e.g., gender or race). This condition would, however, be violated if, for example, students with low pretest scores were more likely to be missing the post-test score because, for example, they refused or were unable to complete the test, or their parents were more likely to fail to provide consent for the outcome testing. MCAR is a strong assumption and one that, in our view, may not be reasonable in most situations.[23]

The second category in Rubin's framework, ***Missing at Random (MAR)***, would hold if the probability of the outcome being missing is unrelated to a student's true test score *after controlling for the other variables in the analysis*.[24] In other words, the missingness is random conditional on the observed X's. For example, the missing data would be MAR if missingness on Y was related to gender but conditional on gender—that is, among boys or among girls—the probability of missing data is a constant. This condition would, however, be violated if, for example, students missing the post-test score were those students who would have scored lower (had they been tested) than those students who were actually tested (those without missing data). It is impossible, of course, to determine if this condition exists because the data on the untested students is missing so one cannot compare the scores for the tested and not tested students.

If the assumptions of MCAR or MAR are true, the missing data mechanism can be considered "ignorable" (MCAR) or correctable (MAR); i.e., in effect the factors that cause missingness are unrelated (or weakly related) to the parameters to be estimated in the analysis. Under such conditions, a variety of techniques discussed in Chapter 3 are available to deal with missing data. In some situations, however, one cannot reasonably assume such ignorability, a category called ***Not Missing at Random (NMAR).*** In these situations, the methods that are available (discussed at the end of Chapter 3) require good information about the determinants of missingness to model the causal mechanism, and not surprisingly, the results of an impact analysis in a NMAR situation are quite sensitive to one's choice of assumptions and statistical approach.

[23] One can see if there are differences by comparing the rates of missing data across categories of the available data but this would not rule out differences between missing and non-missing cases on unobservables.

[24] This is also referred to at times as "covariate dependent missingness" (e.g., Horton & Kleinman, 2007).

3. Selected Techniques for Addressing Missing Data in RCT Impact Analysis

This chapter describes a selected set of techniques that are available to educational researchers to deal with the problem of missing data in group randomized trials.[25] As discussed in Chapter 1, the methods were selected based on a review of several recent articles by experts in the field (Graham, 2009; Schafer & Graham, 2002; Allison, 2002; and Peugh & Enders, 2004)[26] as well as a review of the techniques that have been used in RCTs recently sponsored by the U.S. Department of Education.[27]

As shown in the chart below, some of the methods discussed in this chapter can only be used to address missing data for the dependent or outcome "Y" variable (e.g., student post-test scores), others are only applicable for missing data on the independent "X" variables (e.g., student demographics and pretest score), while some can be used to address missing data problems for both types of variables.

	Can be Used for Missing Data in….	
Methods Discussed	***X Variables***	***Y Variable***
Imputation Methods	√	√
Maximum Likelihood Estimation	√	√
Dummy Variable Adjustment	√	
Weighting Methods[28]		√
"Fully-Specified" Regression Models		√
Selection Modeling		√
Pattern Mixture Modeling		√

The discussion of these different methods is organized into two parts. The first deals with what we refer to as "standard" missing data methods that are in common use, particularly

[25] General issues in conducting group randomized trials, independent of missing data, are covered in many excellent publications. See for example Klar & Donner (2001) from the medical literature, and Bloom (2005) concerning social policy experiments. For a thorough discussion of missing data issues and analysis options in studies that randomize individuals rather than groups see Carpenter & Kenward (2007).

[26] As discussed elsewhere, we intentionally include methods that are commonly criticized in the literature—particularly listwise deletion and simple mean value imputation—for two reasons: the use of these methods is widespread in education (see Peugh & Enders, 2004); and because we are focusing on RCTs as conducted in education, we want to understand how different missing data strategies performed within this unique context, including methods that may have shortcomings in more general applications..

[27] For example, case deletion is commonly used to address missing outcomes (e.g., Bernstein, et al., 2009; Corrin, et al., 2009; Garet, et al., 2008), but studies sometimes use multiple imputation (e.g., Campuzano, et al., 2009) or re-weighting (e.g., Wolf, et al., 2009) to address the missing data problem.

[28] Weighting methods could in theory be used to address both missing X variables and missing Y variables. However, in our experience, RCTs in education never use weighting methods to address missing X variables. In our opinion, this may be because researchers are reluctant to drop observations with missing values of the X variables and re-weight the observed sample members.

when one can assume that missing data are MAR:[29] imputation methods, maximum likelihood estimation, dummy variable adjustment, weighting methods, and fully-specified regression models. The second section focuses on two methods that have been developed to address situations where the missing data can be considered to be NMAR:[30] selection modeling and pattern-mixture modeling. In this second section, we also discuss the use of sensitivity testing that can be used to enhance the reporting of RCT findings under either missing data circumstance.

A. Standard Missing Data Methods[31]

Imputation Methods

Imputation methods handle missing data by "filling in" missing values to create a complete data set for subsequent analysis. For the purposes of discussing imputation methods, we assume that impacts are estimated using a model of the following form:

$$Y_{Post} = \beta_0 + \beta_1 Trt + \beta_2 Y_{\mathrm{Pre}} + \beta_3 x_1 + \beta_4 x_2 + ...\beta_{k+2} x_k + \varepsilon$$

In this ordinary least squares regression model, the impact of the treatment under investigation (*Trt* = 1 if assigned to the treatment group, and =0 if assigned to the control group), on a particular outcome measured post-treatment (Y_{Post}) is estimated controlling for a pre-treatment score on the same measure (Y_{Pre}), and up to *k* baseline covariates measured prior to randomization ($x_1, x_2, ...x_k$). The concepts discussed in this section are not predicated on the assumption that the impact model is an ordinary least squares model, nor is the use of a pre-treatment score as a covariate a necessary component. The model above, however, will serve as a useful example for illustrating how imputation can be used to deal with missing data and the different ways it can be done.

For simplicity, we start our discussion with the assumption that there are missing values only on the variable x_2, and that we have complete responses for all of the other variables in the analysis model. Our goal is to replace the missing values on x_2 with imputed values so that we can proceed with fitting the impact model on all units that were assessed at the time of the post-test (for now, we ignore cases with missing outcomes).

As a first step in the imputation process, one can determine whether any (or all) of the missing values on x_2 can be *logically imputed* (also known as *deductive imputation*). To illustrate, suppose that x_2 is an indicator for whether or not a student is eligible for free or reduced-price lunch, and for a set of newly-enrolled students the values of x_2 were

[29] These methods in principle can also produce unbiased impact estimates in the NMAR situation with the specification of appropriate models for the missing data mechanism. But, as discussed later, knowing what model is appropriate is the difficulty.

[30] While these two methods constitute ways to model missing data, they do not of themselves provide a means of estimating impacts. For that purpose, they have to be combined with other estimation techniques such as maximum likelihood. It is also worth noting that while they were developed to meet the challenges of NMAR data, these models can also be applied to the MAR situation.

[31] Case deletion, which is examined as part of the missing data simulations in the next chapter, is not specifically described here because it simply involves conducting the analyses using only those cases with complete data.

missing because the school had not yet made an eligibility determination. But, suppose that one had conducted a parent interview at baseline that obtained the same information that the schools use to make their eligibility determinations, such as family income and family size. In this instance one would actually have enough information to determine whether or not an individual student is eligible for subsidized school meals. Consequently, because we would be reasonably confident that we could determine, with a sufficient degree of accuracy, whether the student would be eligible, it would be a good strategy to replace as many of the missing values as possible with this logically-imputed information.

When such logical imputations cannot be made, Little & Rubin (2002) identify two general classes of statistical imputation that can be considered, *implicit modeling* and *explicit modeling*. Procedures such as *hot deck* and *cold deck*[32] imputation are commonly used examples of *implicit modeling* methods because there is an underlying model implied by the computational algorithm that relates the data used in the procedure to the generated imputed values (i.e., the model is not explicitly developed or stated). Consequently, Little & Rubin (2002) argue that the implicit nature of the underlying model makes it more difficult to assess whether the model assumptions are reasonable for a particular application.

Alternatively, *explicit modeling* procedures are based on specifically stated models that relate the observed data to the predicted or imputed values. Because we agree that there is value in stating the imputation models explicitly, and because the explicit procedures are no more difficult to implement than the implicit approaches, we focus our remaining discussion on the following three explicit modeling approaches:

- *Mean value imputation* – each missing value is replaced with an imputed value equal to the mean of the observed data. In the current example, a missing value for x_2 would be replaced with the mean of x_2 calculated over all non-missing cases.
- *Non-stochastic regression imputation* – this approach also involves replacing missing data with imputed values but uses predicted values from a regression model.
- *Stochastic regression imputation* – this approach extends the regression imputation by adding a varying component to the predictions so that the imputed values have the same variance as the observed values.

The stochastic regression methods can be implemented either as a *single imputation*, or as a *multiple imputation*. In single imputation, each missing value is replaced with one unique value resulting in the creation of a single rectangular data set. In multiple imputation, each missing value is instead replaced with several separately derived imputed values, typically five to ten (see Rubin, 1987, 1996, and Little & Rubin, 2002), creating multiple analytical data sets. The impact analysis model is then fit to each of the generated data sets, and the overall estimate of treatment impact, and the associated standard error, is derived from combining the results across the multiple data sets.

[32] In "hot deck" procedures, study participants are first grouped into cells based on their similarity on measured characteristics for which data are not missing. Imputed values are then obtained as a random sample from the non-missing values within the cell in which a given missing data case falls. "Cold deck" imputation is a similar procedure but utilizes data from an external source.

Mean Value Imputation

Mean value imputation is essentially a special case of regression imputation (discussed below). Returning to our example, if we fit a simple model of the form,

$$x_2 = \beta_0 + \varepsilon$$

to the observed (non-missing) values of x_2, then the estimate $\hat{\beta}_0$ is the mean of x_2. In this procedure, we then replace all of the missing values of x_2 with the value of $\hat{\beta}_0$. This model assumes that the missing data mechanism, the process that caused there to be missing values on x_2, is independent of the values on all other measured covariates.

In many instances, this assumption may be difficult to justify. In the example below, individuals with high values on x_1 are more likely to have missing values for x_2. Consequently, if we replace the missing values on x_2 with the mean of x_2, we will distort the relationship between x_1 and x_2. Furthermore, if there were a relationship between x_2 and either the outcome (Y) or the treatment indicator (Trt),[33] or both, then this imputation method will also distort the relationships between x_2 and Y or x_2 and Trt. Additionally, mean imputation will cause the estimate of the standard error of the mean of x_2 to be too small. This is because the numerator of the standard error (σ_{x_2}) will be too small (i.e., the imputed values will all have the same value), and the denominator ($\sqrt{n}$) will be too big (i.e., n is the total number of both observed and imputed values, instead of just the number of observed values). However, it is not clear that the standard error that we really care about, i.e. the standard error of the coefficient for the treatment effect, will be consistently underestimated or overestimated.

Higher Probability of Missing x_2 when Value on x_1 is High

x_1	x_2	
540	…	← Fill in missing values of x_2 with mean of x_2
528	…	
510	355	
508	…	
505	340	
498	…	
491	322	
488	310	
483	305	
477	298	
474	295	
472	300	
470	284	
466	276	
465	280	
460	268	

Missing values on x_2 are represented by dots ("…").

[33] Randomization ensures that in expectation, there is no relationship between x_2 and Trt, but in any one particular obtained sample, there may be a correlation in the data between x_2 and Trt.

According to Little & Rubin (2002), an analysis using mean value imputation does not produce consistent estimates of variances, covariances, or standard errors because this simple (albeit often used) imputation method when paired with standard analysis methods does not account for the uncertainty associated with using imputed values in place of observed values.

A variant of simple mean value imputation involves the grouping of units based on their values on some important characteristic, and then using the respective group means as the imputed values to replace missing values for all individuals in the group (this is essentially a non-parametric version of the parameterized regression imputation model described below). For example, groups might be schools, and for a student in school *j* that has a missing value on x_2, one might use the mean of x_2 among students in school *j* as the imputed value that would be substituted in place of the missing value. For a randomized design when an outcome variable is being imputed, mean value substitution should be implemented separately in treatment and control groups.

Non-Stochastic Regression Imputation

Regression imputation extends simple mean imputation by generating predicted values for missing data (e.g., the missing values of x_2) conditional on other measured variables. These variables should include not just those that have scientific relevance to the research question at hand but also "auxiliary variables"[34] that are potential causes or correlates of missingness of x_2 or that correlate with the value of the variable being imputed (x_2).

To implement regression imputation we would fit a model of the form shown below in which x_2 is the dependent variable and all other covariates from the impact model are used as independent or predictor variables.[35] The process is implemented separately in the treatment and control groups (see discussion below), therefore the treatment dummy does not appear in the imputation model.

$$(3) \qquad x_2 = \beta_0 + \beta_1 Y_{\text{Post}} + \beta_2 Y_{\text{Pre}} + \beta_3 x_1 + \beta_4 x_3 + \ldots \beta_{k+1} x_k + \varepsilon$$

After estimating this equation using ordinary least-squares, we then replace missing values with the predicted values from the model where the predicted value $\hat{x}_2$ is obtained as,

$$(4) \qquad \hat{x}_2 = \hat{\beta}_0 + \hat{\beta}_1 Y_{\text{Post}} + \hat{\beta}_2 Y_{\text{Pre}} + \hat{\beta}_3 x_1 + \hat{\beta}_4 x_3 + \ldots \hat{\beta}_{k+1} x_k \,,$$

where the $Y_{Post},$ $Y_{Pre},$ and x_j terms are values from the data for the particular individual involved.

The assumption underlying the use of a regression model to get imputed values is that, for the observations with the missing values, if the true values were known, then the

[34] See Collins, et al. (2001) for a discussion of "auxiliary variables" and their contribution to the strength of the imputation procedure.

[35] Rubin (1996) recommends the inclusion of as many variables as possible in the imputation model. Thus, all other variables from the impact model should be considered the minimally sufficient set, and inclusion of all other possibly relevant variables measured at baseline, potentially including higher order terms and interactions, is advised.

differences between the true values and the regression imputed values would be uncorrelated with the treatment group indicator or with any other variable in the analysis model. That is, the differences are assumed to be pure random error. This approach is expected to fix the previously described problem of distorting the relationships between x_2 and x_1, Trt, Y_{Post}, etc. But, as before, a limitation of this approach, when paired with the use of standard analysis methods on the filled-in data set, is that the method does not take into account the uncertainty associated with using imputed values rather than observed values for missing data cases.

One important feature of the imputation model specified in equation (3) is that it includes the outcome variable (Y) among the predictors of the missing covariate x_2. Surprisingly, this imputation method—and each of the imputation methods described below—is improved if the imputation of X variables takes Y_{post} into account (and MAR holds and the model is correctly specified). It may seem odd to use the outcome variable to predict one or more covariates in the model because it would appear to create circularity when the dependent variable Y is used to impute an explanatory variable such as x_2, and then this connection is reversed to estimate the outcome equation that provides the main finding of the analysis (i.e., the treatment-control difference in outcomes representing the intervention's impact). But, this is exactly what the experts recommend (see, Little & Rubin, 2002; Moons, et al., 2006; and Allison, 2002).

To see the importance of using the outcome variable in imputing baseline covariates, consider a simple analysis model in which the outcome, a post-intervention test score, is specified as a function of randomly assigned treatment status and the student's pre-intervention test score and a set of demographic characteristics. If there is some unobserved variable associated with unusually high or low test score growth over time for certain students (e.g., an especially motivating teacher at a given grade level), then the post-intervention test scores will contain information about this unobserved variable not captured by the pretest scores or demographic characteristics. In this circumstance, omitting the post-intervention test score will lead to omitted variable bias in the imputation model. However, to avoid attenuation of the impact estimate, one should do *separate imputation for the treatment group and control group observations*.[36]

Do biased estimates in the imputation model for covariates translate into biased impact estimates in an RCT? Not necessarily. If the missing data mechanism is the same for the treatment and control groups, the consequences of specification error for biased

[36] In an RCT, the impact findings do not hinge on the relationship between the X variables and Y but rather on the relationship between the treatment indicator variable and Y (i.e., the effect of assignment to the treatment group on the outcome). The treatment indicator is never missing, and X variables measured prior to random assignment cannot correlate with it in expectation unless their imputed values inject correlation. This possibility is avoided by doing separate imputation for the treatment and control group samples. Absent this separation, the impact of the intervention on the outcome—if non-zero—will influence how cases are classified by the imputation procedure. For example, if girls tend to score better than boys on verbal achievement tests and the intervention raises the scores of all students, the positive relationship between test scores and the covariate X = 1 for girls, = 0 for boys will lead more of the cases with missing data for X in the treatment group to be imputed as girls than of cases with missing data for X in the control group. This makes the imputed version of X correlate positively (in expectation) with the treatment indicator, Trt, even though the original version did not (and cannot, having been measured prior to the creation of the Trt variable). Once divided into separate imputation procedures, higher or lower Y values for students with missing sex information *within the treatment group* can only arise because of the underlying girl/boy contrast and not because the intervention had any impact—and similarly for higher or lower Y values within the control group.

imputations will be symmetric for the two samples, making mean outcomes for the two samples following insertion of imputed values equally biased. Moreover, their difference—the basic measure of impact—will not be biased. However, if the missing data mechanism for baseline covariates differs between the two groups biased coefficients in the imputation model can lead to biased impact estimates.[37]

Regression imputation can be used to impute missing *dependent* variables as well as missing covariates. Here again, it is considered good practice to include all of the covariates from the analysis model in the imputation model, plus any other baseline variables that may be associated with missing data (i.e., including auxiliary variables as noted by Collins, et al., 2001). In addition, if the analysis involves multiple dependent variables─for example, separate outcome tests for reading and math achievement─it is a good idea to include all the other dependent variables in the model when imputing any one outcome, to further protect against omitted variable bias in the imputation model.

Single Stochastic Regression Imputation

Stochastic regression imputation involves the addition of "noise" to the imputed value obtained from regression imputation to ensure that the variation among imputed values is the same as the variation among observed values. When the predictive model for x_2 is fit to the data (e.g., equation (4) above), the residuals from the model are saved.[38] Then to obtain an imputed value, a randomly selected residual from the residual file is added to the predicted value from the regression model (i.e., imputed value = $\hat{x}_2$+ a randomly selected residual). In single stochastic regression imputation, each missing value is replaced with a single imputed value resulting in a single rectangular analysis file.

With this method, the imputed values have the same variance as the observed data. However, Little & Rubin (2002) caution that the use of standard analysis techniques on the filled-in data set may result in incorrect standard errors because the method does not take into account the uncertainty associated with using imputed values in place of observed values.

Multiple Stochastic Regression Imputation

Multiple stochastic regression imputation methods have been developed to account for the uncertainty caused by using imputed values. Conceptually the procedure is rather simple. First, imputed values are generated from a stochastic regression procedure like the one illustrated above 5 or 10 times, resulting in separate data sets for each replicated

[37] This problem should be less severe in RCTs than in quasi-experimental studies due to randomization. In a RCT, if the baseline data are collected prior to randomization, the missing data mechanism *cannot* differ for treatment and control groups because the missing data process occurred prior to randomization. However, in RCTs that obtain consent, collect prior year test scores, or conduct "baseline" classroom observations after schools have been randomized (e.g., randomization over the summer with baseline data collection in early fall), then the missing data mechanism for the covariates in the analysis model *could* differ between experimental groups.

[38] Residuals are the difference between the observed values and model predicted values (i.e., the residual for the ith unit is calculated as $r_i = x_{2i} - \hat{x}_{2i})$.

imputation.[39] Assuming that 10 data sets are created, the imputed values across each of the ten data sets are likely to be slightly different from one another because in each instance a randomly selected residual (likely to be different from draw to draw) will be added to the predicted value to obtain the imputation value.

Next, one fits the selected impact model to each of the ten data sets. For example, let "D" be the number of multiple imputations; in the current example D=10. Let "d" be an index for each of the ten data sets (d = 1, 2, … 10). The model for the dth data set is of the form:

$$Y_{Post,d} = \beta_{0,d} + \beta_{1,d} Trt_d + \beta_{2,d} Y_{\text{Pre},d} + \beta_{3,d} x_{1,d} + \beta_{4,d} x_{2,d} + \ldots \beta_{k,d} x_{k,d} + \varepsilon_d$$

In this example, the process would produce 10 estimates for each parameter in the model:

$$\begin{array}{l} \hat{\beta}_{0,1}, \hat{\beta}_{0,2}, \ldots \hat{\beta}_{0,10}, \\ \hat{\beta}_{1,1}, \hat{\beta}_{1,2}, \ldots \hat{\beta}_{1,10}, \\ \ldots, \\ \hat{\beta}_{k,1}, \hat{\beta}_{k,2}, \ldots \hat{\beta}_{k,10}, \ldots \end{array}$$

The final estimate for a particular parameter (e.g., the treatment effect) is the mean over the estimates from the ten repetitions:

$$\overline{\beta}_1 = \frac{\sum_{d=1}^{10} \hat{\beta}_{1,d}}{10}$$

The standard error of the combined estimate ($s.e.\ (\overline{\beta}_1)$) is calculated from (1) a within-imputation variance component, (2) a between-imputation variance component, and (3) an adjustment factor for the number of repetitions (D).[40] Let W_{1d} be the estimated variance of the parameter from repetition d, $(\hat{\beta}_{1,d})$ i.e., $W_{1d} = [s.e.(\hat{\beta}_{1,d})]^2$. Then the **within-imputation variance** is the average of the D=10 estimated variances, calculated as:

$$\overline{W}_{1D} = \frac{\sum_{d=1}^{10} W_{1d}}{10}$$

The standard error estimate of a parameter is provided in all regression modeling software products. Therefore, $\overline{W}_{1D}$ is easily calculated by squaring the standard errors and taking the mean over the ten sets of computer output.

[39] For modest to relatively large amounts of missing data, little is gained by doing more than five to ten repetitions (see Rubin, 1987).

[40] Formulas for the standard error of the combined estimate are from Little & Rubin (2002).

The **between-imputation variance** component is given by the following equation:

$$B_{1D} = \frac{1}{D-1}\sum_{d=1}^{D}(\hat{\beta}_{1,d} - \overline{\beta}_1)^2$$

and the total is given by:

$$T_{1D} = \overline{W}_{1D} + \frac{D+1}{D}B_{1D}$$

In this equation, (D+1)/D is the **adjustment factor for D repetitions**. Thus, the standard error of $\overline{\beta}_1$ is as follows:

$$s.e.(\overline{\beta}_1) = \sqrt{T_{1D}}\ .$$

Once the analyst has calculated $\overline{\beta}_1$ and $s.e.(\overline{\beta}_1)$, the statistical significance of the treatment effect can be calculated in the usual way by comparing the ratio $\frac{\overline{\beta}_1}{s.e.(\overline{\beta}_1)}$ to the quantiles of a t-distribution with v degrees of freedom, where $v = (D-1)(1 + \frac{1}{D+1}\frac{\overline{W}_{1D}}{B_{1D}})^2$. Estimates and standard errors for other model terms are calculated in a similar fashion.

The multiple imputation process described above does not require any specialized software. First, one creates the different data sets required to make the procedure a multiple, rather than a single, stochastic regression imputation. And, then one repeats the impact analysis separately for each of the multiple data sets and calculates the mean of the parameter estimates. The formulas shown above for estimates of the standard errors and degrees of freedom can be programmed in practically any statistical analysis package or even in a simple spreadsheet. Consequently, the process can be done using whatever software the analyst usually uses to do data processing and analysis.

One common question that arises is which covariates should be used when creating a model to obtain predicted values for a covariate that has missing values. The general recommendation of Little & Ragunathan (2004) is to use every available variable in the prediction model−there is no benefit of parsimony in the prediction model. In the case of stochastic regression imputation, one should also use the outcome variable (Y_{Post}) on the right-hand side of the prediction model. While many analysts may be bothered by the seeming circularity of using Y_{Post} in the prediction of imputed values for x_2, and then subsequently using x_2 in the prediction model for Y_{Post}, it is strongly argued in Little & Rubin (2002) and Allison (2002) that omission of the dependent variable (Y_{Post}) from the imputation process can lead to downward bias in regression coefficients,[41] and can lead to inconsistent estimates.[42] This leads to the following guidelines:

- Any variable that will be used in the analytic model (for our discussion the analytic model is the treatment impact model) should also be included in the imputation model, including the treatment status indicator variable; and,
- One should err on the side of using all available information in the imputation model, rather than aiming for parsimony.

For more information on the choice of covariates for the imputation model see the SAS 9 documentation for PROC MI, and other suggested references including Rubin (1996), Barnard & Meng (1999), and van Buuren, Boshuizen, & Knook (1999).

Bayesian Methods for Multiple Imputation

There is one additional source of uncertainty in the parameter estimates that is not properly accounted for in the multiple stochastic regression imputation process described above. As a result, the process described in the previous section is referred to as "improper" multiple imputation. While this procedure largely addresses the problem that multiple imputation was designed to solve—that standard error estimates from single imputation methods are biased downward—the standard error estimates will still be biased downward. This section presents an approach to remove the remaining bias.

Returning to our example, suppose we used an imputation model of the form shown below to model the non-missing values of x_2 (as noted above, the treatment and control groups are modeled separately so there is no treatment dummy in the imputation model):

$$x_2 = \beta_0 + \beta_1 Y_{\text{Post}} + \beta_2 Y_{\text{Pre}} + \beta_3 x_1 + \beta_4 x_3 + ...\beta_{k+1} x_k + \varepsilon$$

We then replace each missing value with an imputed value, $\tilde{x}_2$, where $\tilde{x}_2$ is the sum of the model-predicted value $\hat{x}_2$ and a randomly chosen residual, *r*. Note that the model-predicted value is obtained as shown below where each beta-hat ($\hat{\beta}$) is an estimate from the model above.

$$\hat{x}_2 = \hat{\beta}_0 + \hat{\beta}_1 Y_{\text{Post}} + \hat{\beta}_2 Y_{\text{Pre}} + \hat{\beta}_3 x_1 + \hat{\beta}_4 x_3 + ...\hat{\beta}_{k+1} x_k \ .$$

The extra source of uncertainty arises because the beta coefficients in the prediction model are treated as true parameters instead of sample estimates. Rubin (1987) and Little & Rubin (2002) refer to a multiple imputation process that does not account for the use of sample estimated parameters to get the regression-predicted value as "improper."

The implementation of "proper" multiple imputation adds a layer of complexity to the imputation process. The methods are based on Bayes theory and relate the complete posterior distribution given no missing data to the posterior distribution given observed data. The methods involve forms of Markov Chain Monte Carlo algorithms, and are described by names such as data augmentation (Allison, 2002; Schafer, 1997; Little &

[41] Allison (2002) cites Landerman, Land, & Pieper (1997).

[42] Inconsistent estimates do not get closer to truth as the sample size increases. Little & Rubin (2002) provide citations to more details on inconsistency of estimation if the dependent variable is omitted from the imputation procedure.

Rubin, 2002) and Gibbs sampling (Little & Rubin, 2002; Little & Raghunathan, 2004). Unlike the stochastic regression multiple imputation process described in the previous section, these procedures would not be straightforward for a typical analyst to program using common data management and model fitting software. While there are data augmentation and Gibbs sampling routines available in some standard software packages such as the "Proc MI" procedure in SAS, these are limited to simple imputation models (SAS Institute, 2003).

For example, in many education RCTs where most or all of the analytical models for estimating program impacts will be multilevel models with two or three levels (e.g. students at level 1 nested in schools at level 2), if one uses the improper procedure described above, the imputation model for obtaining predicted values for missing items can be of the same form as the analytic model. That is, the imputation model can be a two-level model with students nested in schools, and the correlation of students within schools can be accounted for in the same manner in the imputation model as it will be in the analytic model for program impacts. Using a procedure such as Proc MI, a researcher would have to assume a simpler multivariate normal model for the imputation model, with no random effects to capture the effects of the clustering of students within school. In this situation, the researcher might, for example, use dummy variables for schools in the imputation model to approximate the multilevel structure that will be used in the analytic model for program impacts.

Analysts who would like to implement multiple imputation approaches to deal with missing data may be faced with what looks like two unsatisfactory options: (1) utilize a "proper" technique that properly propagates the uncertainty caused by having estimated instead of true parameters in the imputation prediction model, but has a mis-match between the imputation model and the impact model; or (2) utilize a multiple imputation technique that is "improper" but allows for an accurate match between the imputation model and the impact model, and may also more easily allow for prediction models for missing variables that are binary, ordered categorical, or unordered categorical. Fortunately, as Allison (2002) and Little & Ragunathan (2004) suggest, in terms of the extent to which methods produce correct standard errors, there is a large jump from single regression imputation to multiple stochastic regression imputation, then a much smaller difference between improper and proper stochastic regression multiple imputation.

Examples of Multiple Stochastic Regression Imputation

In this section we discuss some pragmatic approaches to the implementation of multiple imputation under various conditions that are frequently encountered in impact analyses for educational and other social science outcomes (some additional guidance for multiple imputation is provided in Appendix B).

<u>Missing Values on Binary Variable:</u> When there are missing data on a binary variable (coded 0 or 1), a logistic regression model can be substituted in place of the linear regression model used in the example above. In the example below, we assume that the imputation model is fit separately to treatment and control groups, and we are using both pretest and post-test in obtain predicted probabilities:

$$\log(\frac{\pi_{x_2}}{1-\pi_{x_2}}) = \beta_0 + \beta_1 Y_{\text{Post}} + \beta_2 Y_{\text{Pre}} + \beta_3 x_1 + \beta_4 x_3 + ...\beta_{k+1} x_k$$

where π_{x_2} is the probability that $x_2 = 1$.

Having estimated this model with logistic regression or an equivalent maximum likelihood estimation numeric algorithm, one then calculates the predicted probability that $x_2 = 1$ as:

$$\hat{\pi}_{x_2} = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 Y_{\text{Post}} + \hat{\beta}_2 Y_{\text{Pre}} + \hat{\beta}_3 x_1 + \hat{\beta}_4 x_3 + ...\hat{\beta}_{k+1} x_k)}{1+\exp(\hat{\beta}_0 + \hat{\beta}_1 Y_{\text{Post}} + \hat{\beta}_2 Y_{\text{Pre}} + \hat{\beta}_3 x_1 + \hat{\beta}_4 x_3 + ...\hat{\beta}_{k+1} x_k)}).$$

Next, a random variate is generated from a binomial distribution with probability equal to $\hat{\pi}_{x_2}$. All major statistics software packages have functions to generate random variates in this fashion. This randomly generated value is then used as the desired imputed value. For multiple imputation, if there were to be 10 replications, and if $\hat{\pi}_{x_2}$ were equal to 0.80, one would expect that over ten multiple imputations, about eight of the imputed values would be ones, with the remainder equal to zero.

<u>Missing Values on Multi-category Categorical Variables:</u> If there are missing values on a variable that has more than two categories, and is either ordinal (has natural ordering such as high, medium, and low), or nominal (no natural ordering such as race and ethnicity categories), then ordinal logistic regression, or polytomous logistic regression, can be used to obtain imputed values in a manner similar to that described above for binary variables using logistic regression.

<u>Missing Values on More than One Covariate:</u> Usually, when one is faced with a missing data problem on one variable, the problem also exists for other variables. In the case of a monotone missing pattern like the pattern depicted below, where "." and "x" indicate missing and non-missing values, respectively, the solution is to regress x_2 on x_1 to obtain imputed values for x_2, and then to regress x_3 on x_1 and x_2 (including the imputed values of x_2) to obtain imputed values on x_3, and so on.

Monotone Missing Pattern				
Unit ID	x_1	x_2	x_3	x_4
1	x	x	x	X
2	x	x	x	X
3	x	x	x	.
4	x	x	.	.
5	x	.	.	.

When faced with a general, non-monotone pattern of missing data, one can start with the variable that has the least amount of missing data, and impute values using stochastic regression multiple imputation to fill in the missing values using as much data as is available in the current step.

For example, for the pattern depicted below, one could impute the missing value of x_1 for Unit ID=2, by creating a data set comprised of all individuals with patterns like those shown for Unit IDs 3 and 4, and regressing x_1 on x_2 and x_3. Then, the missing values on x_2 would be imputed by regressing x_2 on x_1, and so on.

A General, Non-Monotone Missing Pattern

Unit ID	x_1	x_2	x_3	x_4
1	x	x	.	X
2	.	x	x	.
3	x	x	x	.
4	x	x	x	.
5	x	.	.	.

An alternative to the process described above is to use Markov Chain Monte Carlo (MCMC) methods to impute enough values to produce a monotone missing data pattern, and then follow with stochastic regression multiple imputation to fill in the remaining cells. The MI Procedure in SAS Version 9 has an option to proceed in this manner under the assumption of a multivariate normal distribution for the data.

Missing Values on a Variable that is Used to Create Subgroups for Subgroup Analysis: Suppose, for example, that a study aim were to estimate separate treatment effects for boys and girls, and that there were missing values on the variable that defines the subgroups (SexMale =1 if boy, =0 if girl). If the subgroup analysis is conducted by using interaction terms in impact analysis model (e.g., SexMale * Trt), then the analyst would multiply impute the missing values on the SexMale variable, and fit the model with the interaction term to the separate data sets. The estimates and standard errors from the multiple imputation model are calculated in the same manner as described previously. So, the fact that the imputed variable defines a subgroup does not present any special problems.

If the approach to the subgroup analysis involves conducting analyses on subsets of data (e.g., making a data set that includes only boys, and estimating the treatment effect using that subset of data), then when the multiply imputed variable SexMale is used to create separate data subsets, the data sets vary slightly over the replications because of the stochastic nature of the imputation for SexMale. Again, each of the replications will produce a separate estimate, but they can be combined as described previously.

Maximum Likelihood (ML) Estimation

Maximum Likelihood (ML) is a statistical estimation method for identifying population parameter values that can be used in a variety of ways to deal with missing data. The different methods, described in more detail below, are:

- Separate estimation of conditional and marginal distribution functions using all available data in each instance, then solving for the impact estimate as a function of the distributions' parameters;
- Prediction of missing values using the expectation maximization (EM) algorithm, with impact estimation conducted on a data set that replaces missing values with predicted values;

- Prediction of missing values using the EM algorithm multiple times, with a random residual added and impact estimation conducted multiple times and the separate results averaged;
- Production of sufficient statistics using the EM algorithm, from which the impact model can be estimated; and,
- Full information maximum likelihood that maximizes the joint distribution function of all the data, missing and non-missing.

In general, ML estimation involves (a) specifying a "likelihood function"—i.e., the probability distribution function assumed to generate the data, (b) substituting the data into this function, and (c) finding the parameter values for the distribution that maximize the likelihood function. In other words, ML seeks the parameter values that make the observed data as likely as possible to have occurred. For some likelihood functions, there is a closed form solution to the maximization problem, and the parameter estimates that maximize the likelihood function can be identified using differential calculus. In most applications, ML estimates are identified through an iterative process.[43] This technique can be applied to situations in which none of the data are missing, but it can also be applied to situations with missing data.

To illustrate how ML estimation works with complete data, consider the type of linear model we typically specify in an RCT:

$$Y = \beta_0 + \beta_1 Trt + \beta_2 x_1 + \beta_3 x_2 + ...\beta_{k+1} x_k + \varepsilon$$

where, Y is the student outcome variable, Trt is a variable that indicates the group to which the student was randomly assigned (Trt = 1 if treatment group, =0 if control), and X is a vector of k covariates measured prior to randomization. ML estimation requires the analyst to make an assumption about the joint distribution of all these variables. Commonly, we assume that the variables have a joint normal distribution. If some of the variables are not normally distributed (e.g., dichotomous outcome variables), other models can be specified and estimated via ML methods (e.g., probit and logit models).[44] ML estimation chooses values for all the parameters of the joint distribution of Y, Trt, and the x variables, including the impact estimate β_1. If the joint normality assumption holds, ordinary least-squares (OLS) can be used to compute ML estimates, i.e., OLS is equivalent to maximum likelihood estimation under the joint normality assumption.

When some of the data are missing, ML can be extended in a number of ways, each of which are reviewed below with their associated strengths and weaknesses.

[43] The iterative process involves making small changes to the parameter estimates and re-evaluating the likelihood function. The process ends when further small changes to the parameter estimates leave the likelihood function virtually unchanged, indicating that a "peak" or maximum point has been attained.

[44] Moreover, as Allison (2002) asserts: "Although the assumption of multivariate normality is a strong one, it is completely innocuous for those variables with no missing data" (p. 18), such as is true for the treatment status indicator, Trt.

(1) Separate estimation of conditional and marginal distribution functions

The first way to treat missing data using ML methods is to express the joint distribution function to be maximized as the product of a conditional distribution and a marginal distribution:

$$\mathrm{f}(Y, Trt, X) = \mathrm{f}(Y \mid Trt, X)\, \mathrm{f}(Trt, X)$$

With missing data, ML can be used to estimate the parameters associated with each of these two distributions *separately*[45] using only the observations for which the relevant variables are observed. For example, suppose only the outcome variable Y was missing for some cases, and X and Trt were never missing. This technique would (a) estimate the parameters of the conditional distribution function $\mathrm{f}(Y \mid Trt, X)$ using only those cases with non-missing outcomes, and (b) estimate the parameters of the marginal distribution function $\mathrm{f}(Trt, X)$ using all cases. In this way, ML uses all of the available information for cases with missing outcomes and cases with non-missing outcomes. The conditional and marginal distribution functions, with their parameters now estimated using ML methods, are then recombined to get the desired joint distribution, $\mathrm{f}(Y, Trt, X)$. The relationship of Trt to Y—the impact estimate—can be derived from this joint distribution and is unbiased if the right distributional type is adopted for $\mathrm{f}(Y, Trt, X)$ and its components. Moreover, the common assumption of a multivariate normal distribution produces impact estimates that become unbiased as sample size increases, regardless of the true distributional form. As Allison (2002) states, "Maximum likelihood under the multivariate normal assumption produces consistent estimates of the means and the covariance matrix [from which all distribution parameters can be derived] *for any [true] multivariate distribution* with finite fourth moments" (p. 88, emphasis added).

This approach to ML with missing data is only feasible when the missing data are hierarchical—i.e., when missingness of a particular variable implies missingness for all variables with higher overall missing data rates. For a joint distribution involving three variables, this condition is satisfied if—in the case where Y is most often missing followed by x_2 followed by x_1—all observations with x_1 missing are also missing x_2 and Y, and all observations with x_2 missing are also missing Y. A dataset can be made hierarchical in its missing data patterns by dropping observations that violate the pattern. However, this would sacrifice useful information. Alternatively, analysts can impute values for the less commonly missing variables in order to achieve a hierarchical dataset. When conducting an impact analysis in an RCT, one can sometimes achieve a hierarchical missing data pattern with minimal imputation if the missing data rates for pre-randomization covariates are low. In these situations, one need only impute values for the X variables where x is missing but Y is non-missing to create a hierarchical dataset.[46] This has the disadvantage, though, of requiring two different procedures, imputation followed by ML estimation, with the first of the two not defined by the method.

[45] If each distribution is maximized, so is their product.

[46] The treatment/control indicator variable *Trt* is never missing (because the experiment creates it).

(2) Prediction of missing values using the expectation maximization (EM) algorithm

Unlike option (1), the other ML methods on the list above do not require hierarchical missing data. Several of these use the expectation-maximization (EM) algorithm discussed by Allison (2002, p. 19) and Graham (2009, p. 6.7). EM sets initial parameter values for the chosen distributional type (usually normal) and then uses those values to predict values for missing variables. It then puts all the data, real and imputed, into the likelihood function and solves for *updated* parameter values that maximize the likelihood function. The new parameters are used to re-impute the missing values, and the process iterates between ML parameter calculation and missing data imputation until the value of the likelihood function stops increasing (i.e., is maximized). This gives the final set of imputed values for use—along with the non-missing data—in the impact regression model.

As noted by both Allison and Graham, the EM algorithm when used in this way may produce incorrect standard errors for regression coefficients, including the impact estimate. Indeed, Graham (p. 6.8) says that "Standard errors…will be too small, sometimes to a substantial extent."

(3) Prediction of missing values using the EM algorithm multiple times

Graham (p. 6.8) offers two ways to obtain unbiased standard error estimates when using ML in the presence of missing data: direct maximization of the likelihood function, inclusive of missing values (see (5) below), and multiple imputation using the EM algorithm to impute values. Multiple imputation using EM parallels the multiple stochastic regression imputation method discussed earlier, except that it uses the EM algorithm, rather than a regression model, to obtain each set of imputed values. In both cases, a random residual is added to the imputed values to ensure unbiased standard error calculations. When run multiple times, this method produces N impact estimates and N standard errors, one for each EM-generated set of imputed values that replaces missing data. The N sets of findings are then combined as in multiple stochastic regression imputation to give an overall impact estimate and its correct standard error. Computer programming for this procedure is very similar to multiple stochastic regression imputation.

(4) Create "sufficient statistics" using the EM algorithm

Rather than impute values for missing cases, the EM algorithm can be used to produce estimates of the means and covariances of all the variables in a dataset, then these "sufficient statistics" can be used to estimate an impact model. However, the use of sufficient statistics rules out the two-level random intercepts impact model that is typically used in education RCTs applications.

(5) Full information maximum likelihood (FIML)

The final ML approach from the literature—full information maximum likelihood (also called direct maximum likelihood)—maximizes the joint distribution function of all the data, missing and non-missing, in a single step. This method is recommended by many experts (e.g., see Allison (2002) and Graham (2008)), who indicate that it yields regression coefficients and standard error estimates that are approximately unbiased. At the same time, full information maximum likelihood is quite difficult to implement as Graham notes (p. 6.10) "FIML methods deal with the missing data, do parameter estimation, and estimate standard errors all in a single step. This means that the regular, complete-cases algorithms must be completely rewritten to handle missing data." However, there are specialized software packages written for structural equations modeling (e.g., Mplus, AMOS) that can be used to implement these algorithms.

Based on this review of the different methods of maximum likelihood estimation that are available in the literature, multiple imputation using the EM algorithm is in our opinion the method most useful to test in the simulation analysis discussed in Chapter 4. It is the one variant that (a) can be implemented with all types of missing data (not just hierarchical missing data), (b) gives correct standard errors for impact estimates, (c) allows estimation of the two-level random intercepts impact model frequently used in educational RCTs, and (d) does not require specialized computer software or expertise.

Dummy Variable Adjustment for Missing Covariates

A simple alternative to the imputation of missing covariates is a method commonly called "dummy variable adjustment" that involves three related steps:

1. ***Create a new variable Z*** − Z is set equal to X for all cases where X is non-missing and set to a constant value, C, for those cases where X is missing. C is often set to 0 or the mean of X, but it does not matter which value is used.
2. ***Create a new variable D*** – This dichotomous variable is set equal to one for those cases where X is missing, and set equal to zero for those cases when X is not missing.
3. ***Replace X in the impact analysis model with Z and D*** – With this new specification the impact model will estimate the relationship between Y and X when X is not missing, and it will estimate the relationship between Y and D when X is missing.

For example, if there were 100 observations in a data set, and X is missing for 15 of the 100 observations, we would: (1) set Z to zero and D to one for the 15 missing observations; (2) let Z equal X and set D to zero for the other 85 non-missing observations; and, (3) include both Z and D in the model in place of X.

The academic literature seriously questions this approach for dealing with the general case of missing data. For example, Jones (1996) shows that in the general case, this approach leads to biased estimates of the coefficients in the regression model, and Allison (2002) shows a numeric example that lends support for this argument. It is easy to imagine that in general, the dummy variable approach would produce biased estimates of the relationship between Y and X controlling for other variables.

However, although the dummy variable adjustment might not work well in general, we believe there are reasons why it may work well in the special case of an RCT. In RCTs with complete data, inclusion of the right covariates in a correctly specified impact model can help increase the precision of the impact estimates. *However, these covariates are not necessary to obtain unbiased impact estimates.* If the dummy variable adjustment's main drawback is a misspecification of the functional form of the analysis model, this may not lead to bias in the impact estimates in an RCT. In fact, Jones (1996) shows that that the dummy variable adjustment will generally produce biased impact estimates, but that the impact estimates will be unbiased if assignment to treatment is uncorrelated with the covariate that has some missing data. Because random assignment ensures that assignment to treatment is, in expectation, uncorrelated with all observed covariates, as well as unobservables, it would seem that the general concerns about the dummy variable method may not apply in RCTs.

Of course, in real education studies, the covariates may not always be uncorrelated with assignment to treatment. For example, consider an RCT which randomizes schools over the summer and collects baseline data in the fall when school resumes. The level of cooperation from school staff in obtaining completed consent forms could be higher in treatment schools than in control schools. Furthermore, willingness of parents to sign a consent form may be associated with whether a school received the intervention or not. Therefore, in real education RCTs, the covariates—even the dummy variable itself (in evaluations that use the dummy variable adjustment)—could be correlated with assignment to treatment.

Therefore, it is not clear whether one should expect the dummy variable adjustment to perform well or poorly in reducing bias in real RCTs in education. In light of how often this technique is used, we believe it deserves additional scrutiny and empirical testing to determine whether it should or should not be used in RCTs in education.

Re-weighting Methods

Another commonly used approach to account for missing data in the outcome variables involves re-weighting the observed data. To understand the problem that re-weighting is designed to address, consider the example of a follow-up survey of students with a response rate of 75 percent. If the probability of responding to the survey—and providing data for survey-based outcome variables—were the same for all students, there would be no problem other than a reduction in sample size. A 75-percent sample is smaller and provides less statistical power than a 100-percent sample, but the loss of 25 percent of the initial sample would not necessarily introduce any bias.

However, it is well known that certain groups are more likely to respond to surveys than others and individuals from these groups will be overrepresented among survey respondents (Little, 1986; Lessler & Kalsbeek, 1992). Therefore, it is quite possible that in our example, the response rate could be higher than 75 percent for some groups and lower than 75 percent in other groups. This variation in response patterns can skew the sample of individuals for whom there are complete data towards those who are more likely to respond to surveys. The same logic applies to student-level administrative data from schools.[47]

Skewing of the sample is likely to create a problem in most RCTs because it may lead to non-response bias in the average outcomes for the treatment and control groups. If the bias is the same in both groups for each outcome measure, then the impact estimates will

[47] If follow-up school records data for students are less likely to be available for students who move outside the school system, and some types of students are more likely to leave the schools system than others—perhaps for private schools, independent charter schools, or public schools in other districts—then the sample of students for which follow-up student data are available may be skewed toward the types of students who are likely to remain in the district's school system.

be unbiased. However, if the missing data mechanism is different for the two groups, the bias may be different as well, and this leads to biased impact estimates.

Re-weighting deals with the problem of non-response bias by assigning larger weights to groups that are underrepresented among survey respondents—those with lower than average response rates—than to other groups that are overrepresented among survey respondents. Put differently, we can "weight down" respondents from groups with high response rates and "weight up" respondents from groups with low response rates. This approach can ensure that the weighted distribution of survey respondents matches the distribution for the complete population of interest for any observed variable, including those that may be systematically related to whether or not the sample member completes the follow-up survey.

It is easy to see how re-weighting works in settings where we only have one group—maybe program participants—and we want to measure an average outcome for this group. For example, suppose our sample is split evenly between boys and girls, but girls are less likely to take a math achievement test than boys—say 1/3rd of girls and 2/3rds of boys take the test. In this scenario, *the sample with available test data* is skewed toward boys: while the ratio of boys to girls in the overall sample is 1:1, the ratio among the students with data is 2:1. If girls score higher than boys on the test, the skewing of the sample toward boys will depress the average post-test score and understate the average test score for the sample as a whole; if boys score higher than girls on the test, the skewing of the sample toward boys will have the opposite effect.

To address this problem, we could simply re-weight students by the inverse of the response rate for the group to which the student belongs. Girls would receive a weight of $(1/3)^{-1}$ or 3, and boys would receive a weight of $(2/3)^{-1}$ or 1.5. Because every girl with outcome data receives twice the weight of every boy with outcome data, the weighted sample of test data is no longer skewed toward boys. This approach extends naturally to RCTs with two or more experimental groups—such as one treatment group and one control group—but the response rates should be calculated and weights computed separately for each group.

Re-weighting can eliminate non-response bias *if the characteristics used to stratify the sample fully explain the variability in response rates across the sample*. In our example above, constructing separate weights for boys and girls would entirely remove the bias in mean outcomes if there is no variability in the response probability within group (e.g., all boys have the same response probability and all girls have the same response probability).

In constructing weights, a common approach is to use the available covariates to construct "weighting classes", which is similar to post-stratification (Lohr, 1999, p. 268). For example, if the researchers had reason to believe that the probability of responding to the survey varied by race and sex, sex-by-race weighting classes could be constructed, and respondents could be weighted by the inverse of the response rate within their class. This approach will produce unbiased impact estimates of the regression parameters if the missing data are MAR (Lohr, 1999, p. 265). More precisely, this approach yields unbiased parameter estimates under the following condition: "Respondents in weighting class j are a random sample of the sampled units (that is, the data are MCAR within adjustment class j)" (Little & Rubin, 2002, p. 47).

If some of the factors that influence both the probability of responding to the survey and the outcome of interest are not captured by the weighting classes, we would not expect re-weighting to yield unbiased regression coefficient estimates. However, as long as the weighting class variables capture some of these key factors, we would expect re-weighting to reduce the non-response bias in the parameter estimates, including the impact estimate in RCTs.

In constructing weights, researchers are often concerned about the tradeoff between bias and variance (Cox, 1991). As the number of key factors included in creating weighting classes increases, the amount of non-response bias that remains will generally decrease. However, this also increases the number of weighting classes, decreases the sample size in each weighting class, and increases the sampling error or "noise" in the estimated response probabilities. Random noise in the estimated response probabilities will produce additional variability in the weights and increase the standard error of the estimate. Therefore, adding additional weighting classes can reduce bias, but it can also reduce the precision of the impact estimates.

These concerns sometimes lead researchers to conduct tests to determine which of the variables to include in post-stratifying the sample—either directly, as covariates in a propensity score method, or via some other approach—and constructing weights.[48] A common approach is to run a regression of whether or not the sample member responded to the survey as a function of a broad set of variables that may—or may not—help to explain survey response. The process that statisticians use to select variables may vary, but the use of a *p*-value based criterion is common. The final model identifies the observed characteristics that the model suggests are important predictors of survey non-response. These characteristics can then be used to post-stratify the sample for computing response rates and constructing weights.

Unfortunately, the literature as a whole provides little guidance on how to balance bias and variance. In our view, if the increase in variance from including relatively unimportant variables is small, and the risks from excluding potentially important variables are large, it might make sense to exclude only variables with very high p-values (e.,g. >0.50).

"Propensity score matching" provides a particularly sophisticated method for accounting for respondent and nonrespondent differences on all measured characteristics at once by using all available background variables to predict the probability of a particular person's outcome observation being observed. Cases with observed *Y* values in the same predicted propensity score range as cases with missing *Ys* are then re-weighted upward to represent the missing cases as well.[49] Because the predictive model uses all available background variables to compute the predicted "propensity scores," Rosenbaum & Rubin (1984) prove that stratification on the resulting scores does as much to reduce non-response bias

[48] See, for example, Battaglia, et al. (2008).

[49] An alternative approach to using propensity scores for ranges is to form an analysis weight for each observation equal to its inverse propensity score. Baker, et al. (2006) discuss both procedures, and in particular the advantages of the range-based weighting approach over individualized weights.

as would full stratification on the cross-tab of all the individual background variables.[50] Note that in both approaches one must estimate the probability-of-missingness equation separately in the treatment and control groups to ensure that the procedure addresses the possibility of different missing data patterns in the two samples. This method does not addresses selection on unobservables, obviously, nor does it differ fundamentally from other stratified reweighting methodologies for dealing with missing outcome data except through its efficient inclusion of information on all potential stratifiers.

One advantage of the weighting approach is that for users accustomed to conducting analysis with survey weights, also accounting for non-response may be an easy extension of the use of sampling weights. Furthermore, while this is not entirely clear, it may be true that the same procedures used to obtain correct standard errors when survey data need weights to offset differential initial sample-inclusion probabilities[51] still provide correct standard errors when the weights have been adjusted as well for non-response.[52] Finally, if the missing data problem is largely one of "unit nonresponse," such as missing all survey variables for students who did not complete the survey, instead of "item nonresponse," which can generate a scattershot pattern of missing data across different variables, we need to construct only one weight variable per data collection instrument (e.g., survey or school records collection) to address the missing data problem.

On the downside, weights are a cumbersome and little used approach to addressing item nonresponse because it would require constructing a separate weight for each variable with missing data. In contrast, multiple imputation can be used to directly address both unit and item nonresponse. In addition, while more analysts have used re-weighting methods than multiple imputation methods, in our view re-weighting requires as many steps and as much researcher judgment as multiple imputation. Finally, some experts have argued that multiple imputation yields more precise estimates than re-weighting (e.g., Schafer & Graham, 2002).

[50] While it is often argued that one must correctly specify the propensity-to-be-missing equation for this to be true, Baker, et al. (2006) explain that by using the propensity score to stratify before re-weighting (as opposed to giving each individual observation distinct weight equal to the inverse of its own propensity score) this need is eliminated. In particular, they note that "Although the preliminary phase of computing the propensity-to-be-missing score requires the appropriate covariates for modeling the missing-data mechanism, the exact function form is not critical. . . . The reason is that within each [stratum] the probability of missing the outcome is similar for all subjects regardless of the [specification of the propensity score] model." In other words, the computed propensity scores for individual observations do not need to be exact—and hence that equation does not need to assume the correct functional form—for the great majority of observations to be classified into the correct stratum.

[51] In an RCT with unequal sampling probabilities or assignment rates, the weight should be constructed to equal the inverse of the product of the random assignment and response probabilities.

[52] The difference in the case of weighting to offset differential non-response probabilities is that these probabilities are only estimated, not known from the sampling and random assignment procedures themselves. The added uncertainty from estimating rather than knowing the true probability of non-response for different subpopulations may imply that appreciably different (i.e., larger) standard errors are needed than conventional weighted data analysis procedures compute.

Fully-Specified Regression Models with Treatment/Covariate Interactions

Another way to use pre-randomization covariates to adjust for missing outcome data is to interact the covariates in the impact regression with the treatment indicator. Though not as familiar in the literature as the other methods discussed in this report, this approach has appeared in the applied RCT literature (see Bell & Orr, 1994) and builds on the familiar notion of using covariates to adjust for *chance* differences in treatment versus control group outcome levels in impact regressions. Rather than chance differences arising during random assignment, in this application we are dealing with differences between the treatment and control group <u>outcome</u> data arising through differential non-response in the two samples.[53]

To illustrate this approach, let's return to our earlier example:

$$Y = \beta_0 + \beta_1 Trt + \beta_2 x_1 + \beta_3 x_2 + ...\beta_{k+1} x_k + \varepsilon$$

where, *Y* is a student outcome variable, *Trt* is an indicator variable for randomization of the student or his/her classroom or school into the treatment group (*Trt* = 1 if treatment group, = 0 if control), and $x_1, x_2, ...x_k$ are a set of up to *k* covariates measured prior to randomization. Adding terms to the model that interact *Trt* with all of the *x*'s (covariates) yields the following equation:

$$Y = \beta_0 + \beta_1 Trt + \beta_2 x_1 + \beta_3 x_2 + ...\beta_{k+1} x_k + \beta_{k+2}(Trt * x_1) + \beta_{k+3}(Trt * x_2) + ...\beta_{k+k+1}(Trt * x_k) + \varepsilon$$

Once this equation is estimated using ordinary least-squares regression, each individual sample member's impact, *m_i*, can be approximated by subtracting the predicted outcome if the individual were in the control group from the predicted outcome if the individual were in the treatment group:

Trt=1 $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 * 1 + \hat{\beta}_2 x_{1i} + \hat{\beta}_3 x_{2i} + ...\hat{\beta}_{k+1} x_{ki} + \hat{\beta}_{k+2}(1 * x_{1i}) + \hat{\beta}_{k+3}(1 * x_{2i}) + ...\hat{\beta}_{k+k+1}(1 * x_{ki})$

Trt=0 $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 * 0 + \hat{\beta}_2 x_{1i} + \hat{\beta}_3 x_{2i} + ...\hat{\beta}_{k+1} x_{ki} + \hat{\beta}_{k+2}(0 * x_{1i}) + \hat{\beta}_{k+3}(0 * x_{2i}) + ...\hat{\beta}_{k+k+1}(0 * x_{ki})$

Difference $m_i = \hat{\beta}_1 * 1 + \hat{\beta}_{k+2}(1 * x_{1i}) + \hat{\beta}_{k+3}(1 * x_{2i}) + ...\hat{\beta}_{k+k+1}(1 * x_{ki})$

[53] If this were the only difficulty—that cases with non-missing outcome values differ between the treatment and control groups in measured background characteristics—inclusion of those characteristics as covariates in the impact model would be enough to offset the problem assuming the functional form relating *x*'s to *Y* is correct and the magnitude of impact does not vary with the *x*'s, *without* interacting the covariates with the treatment indicator variable. However, we expect that even with differential non-response removed from the data in this way, a base level of non-response common to both the treatment and control groups would remain and could bias impact estimates not because of treatment/control mismatches among cases with observed outcomes but because neither of these samples represents the full population of interest: all units randomly assigned. Thus, while including covariates without interactions can achieve internal validity for measuring impacts on respondents it cannot create the desired external validity in representing the full population of interest. The use of interaction terms described here makes this further adjustment by modeling how impact magnitude varies with background factors and then extrapolating to the background characteristics of the entire experimental sample, including cases whose outcome data are universally missing from both the treatment and control group respondent samples.

The mean impact for all members of the treatment group, $\overline{m}$, is just the average of the m$_i$ calculated for all treatment group members for whom all of the *x*'s (covariates) are fully observed (and hence *m$_i$* can be calculated).[54] The statistical precision of this estimate can be increased by first calculating, and then averaging, *m$_i$* for *all* members of the dataset for whom all of the *x*'s are observed, including members of the control group, since by random assignment these individuals represent the same population as treatment group cases as concerns the *x* variables and provide a major boost to the size of the sample across which *m* is averaged. Thus,

$$\begin{aligned}
\overline{m} \quad &= average(m_i) = \\
average[\hat{\beta}_1 * 1 + \quad &\hat{\beta}_{k+2}(1 * x_{1i}) + \hat{\beta}_{k+3}(1 * x_{2i}) + ...\hat{\beta}_{k+k+1}(1 * x_{ki})] \\
&= \hat{\beta}_1 + average[\ \hat{\beta}_{k+2}(1 * x_{1i}) + \hat{\beta}_{k+3}(1 * x_{2i}) + ...\hat{\beta}_{k+k+1}(1 * x_{ki})] \\
&= \hat{\beta}_1 + \hat{\beta}_{k+2} * average(\ x_{1i}) + \hat{\beta}_{k+3} * average(\ x_{2i}) + ...\hat{\beta}_{k+k+1} * average(\ x_{ki}) \\
&= \hat{\beta}_1 + \hat{\beta}_{k+2}\overline{x}_1 + \hat{\beta}_{k+3}\overline{x}_2 + ...\hat{\beta}_{k+k+1}\overline{x}_k
\end{aligned}$$

where $\overline{x}_1, \overline{x}_2, ...\overline{x}_k$ are the average of the covariates across all cases, treatment and control, where the *x* variables are observed (regardless of whether the outcome *Y* is observed).

This improved impact measure, $\overline{m} = \hat{\beta}_1 + \hat{\beta}_{k+2}\overline{x}_1 + \hat{\beta}_{k+3}\overline{x}_2 + ...\hat{\beta}_{k+k+1}\overline{x}_k$, differs from the original estimate of impact, $\hat{\beta}_1$, the coefficient on *Trt*, to the extent that the magnitude of impact is sensitive to individual sample member characteristics—i.e., to the extent that the coefficients of the interaction terms $(Trt * x_1), (Trt * x_2),...(Trt * x_k)$ differ from 0. It is in this sense that the methodology offsets the fact that the sample for which

$$Y = \beta_0 + \beta_1 Trt + \beta_2 x_1 + \beta_3 x_2 + ...\beta_{k+1} x_k + \beta_{k+2}(Trt * x_1) + \beta_{k+3}(Trt * x_2) + ...\beta_{k+k+1}(Trt * x_k) + \varepsilon$$

can be estimated may not match in its background characteristics the full sample for which *x*'s are observed. Regression without the interactions, calculates the treatment estimate ($\hat{\beta}_1$) as the reflection of treatment-control outcome differences in the restricted sample where *Y* is available. The new specification extends this estimate by projecting it to treatment-control outcome differences *for all sample members where all x's are observed, including those for whom y is not available.*

Of course, the projection is only an approximation, and only as good as the assumptions that underlie it. The impact estimate produced is unbiased if (i) the background *X* variables included in the model encompass all the predictors of non-response, (ii) the relationship of the included background characteristics to untreated outcomes in the control group is linear, and (iii) the relationship of the included background characteristics to the magnitude of impact is linear. Moreover, the latter two relationships need to be the same—i.e., to have the same sets of coefficients—in the

[54] This limits the technique to the subset of the population for which the covariates in *X* are observed, a property common to all missing data procedures that use background variables to adjust impact estimates for potential non-randomness of outcome variable missingness conditional on the *x*'s, such as re-weighting or stratified imputation. This limitation applies unless missing *x* variables are imputed using procedures discussed earlier, in which case the current procedure—like all others that use imputed Xs—encompasses the full sample but is somewhat sensitive to the reliability of the covariate imputation method used.

sample on which the model is estimated (cases with non-missing *Y*) as in the sample as a whole, to which the projection attempts to generalize.

A more sophisticated version of the methodology that allows for non-linear relationships between background characteristics and impact magnitudes includes higher-order interaction terms, such as in the following specification

$$\begin{aligned} Y = \beta_0 + \beta_1 Trt + \beta_2 x_1 + \beta_3 x_2 + ...\beta_{k+1} x_k + \\ \beta_{k+2}(Trt * x_1) + \beta_{k+3}(Trt * x_2) + ...\beta_{k+k+1}(Trt * x_k) + \\ \beta_{k+k+2}(Trt * x_1^2) + \beta_{k+k+3}(Trt * x_2^2) + ...\beta_{k+k+k+1}(Trt * x_k^2) + \varepsilon \end{aligned}$$

The implied "response function" in this equation—the expression conveying the magnitude of the intervention's impact for a student having a particular background profile of *x*'s— is a quadratic rather than simply a linear function of the *x*'s, and hence much more capable of picking up that part of the intervention's effect that becomes progressively more sensitive to the *x*'s or less sensitive to the *x*'s as the *x*'s move further from it their means. Of course, this way of expanding the model's flexibility uses up more degrees of freedom than the simple linear response function.

For any version of the model, one can simplify computation of both the impact estimate and its standard error by differencing the covariates from their observed means in the sample before estimating the equation. In the linear interaction case

$$\begin{aligned} Y = \beta_0 + \beta_1 Trt + \beta_2 x_1 + \beta_3 x_2 + ...\beta_{k+1} x_k + \beta_{k+2}(Trt * x_1) + \\ \beta_{k+3}(Trt * x_2) + ...\beta_{k+k+1}(Trt * x_k) + \varepsilon \qquad [Eq. 1] \end{aligned}$$

can be restated and fit to the data in the form

$$\begin{aligned} Y = \alpha_0 + \alpha_1 Trt + \beta_2(x_{1i} - \overline{x}_1) \quad + \quad \beta_3(x_{2i} - \overline{x}_2) \quad + \quad ...\beta_{k+1}(x_{ki} - \overline{x}_k) + \\ \beta_{k+2}(Trt * (x_{1i} - \overline{x}_1)) + \beta_{k+3}(Trt * (x_{2i} - \overline{x}_2)) + ...\beta_{k+k+1}(Trt * (x_{ki} - \overline{x}_k)) + \varepsilon \quad [Eq. 2] \end{aligned}$$

where α_0 in Eq. 2 is equivalent to $\beta_0 + \beta_2\overline{x}_1 + \beta_3\overline{x}_2 + ... + \beta_{k+1}\overline{x}_k$ from Eq. 1, and α_1 is equivalent to $\beta_1 + \beta_{k+2}\overline{x}_1 + \beta_{k+3}\overline{x}_2 + ... + \beta_{k+k+1}\overline{x}_k$. If, for example, n=1,000 students were randomized to treatment or control, and the pre-treatment covariate values on the *x*'s were known for all 1,000 students, but post treatment outcome measurements (*Y*) were obtained for only 800 students, the model above would be fit to the n=800 observations with observed *Y*'s. But the values of $\overline{x}_1, \overline{x}_2, ...\overline{x}_k$ for those 800 observations would be calculated from all n=1,000 students.

With this model, the coefficient $\hat{\alpha}_1$ is the estimated treatment effect when all covariate *x*'s are at their mean values. This coefficient is identical to $\overline{m}$, described earlier. Thus, the desired impact estimate, $\overline{m}$, and its standard error are easily obtained from this specification, already adjusted for differences in background characteristics between the students whose *Y* outcome measures are observed and those for whom they are not. The standard error is unbiased if observations are independent and homoskedastic (i.e., have equal variance)—the usual assumptions in ordinary least-squares regression. If the data are known to have a different error structure, a different estimation method will be needed for the standard error (but not for the impact estimate)

B. Methods to Address Missing Data that are NMAR

Missingness in the outcome variable, Y, may be associated not just with which students are in subpopulations traced by the background X variables, but also with the true values of Y itself within any subpopulation defined by measured pre-random assignment characteristics of the sample. The potential for biased impact estimates arises in this NMAR situation under all of the methodologies discussed up to this point. Consequently, this section examines three additional methods found in the missing data literature: selection modeling, pattern-mixture models, and the use of bounds and sensitivity analysis. The first two of these methods are ways of modeling the pattern of missingness in the data, and are not impact estimation techniques in their own right; consequently, they are often used with maximum likelihood methods to estimate treatment effects. The third method discussed here, sensitivity analysis, can be used to indicate how far the true impact may lie above or below *measured* impacts.

Selection Modeling

The simplest approach to addressing missing data is to drop the cases with missing values (i.e., case deletion); as indicated earlier, whether this approach introduces bias depends on the missing data mechanism. Selection modeling attempts to model the mechanism by which outcome values become missing for some observations but not others. Correct specification of the selection mechanism behind missingness removes the threat of biased impact estimates. Just as correct model specification can remove selection bias in quasi-experimental studies, correct modeling of the missing data mechanism can remove sample selection bias due to missing data in RCTs.

However, specifying the missing data mechanism correctly is at best difficult; "Given a model for the data…there are infinitely many different non-ignorable missing data mechanisms" (Allison 2002, p. 77). Therefore, because there are always multiple missing data mechanisms that are consistent with the observed data, it is impossible to determine whether the missing data model specified by the researcher is correct or incorrect.[55]

The best-known version of selection modeling is the selection correction model of James Heckman (1976). This model assumes that missingness of the outcome variable Y is triggered by a non-observed "latent" variable related to Y, which we can call L. When L exceeds a particular threshold value, h, Y is not observed. For example, if L were a measure of parent protectiveness, parents may refuse to give consent for interviewing their child if L > h. If L is also correlated with the outcome Y, we have a non-ignorable missing data problem.

In the Heckman procedure, the probability of missing Y data is assumed to follow a probit model, as it would if Y and L are jointly normally distributed. The likelihood function for missing cases follows from this assumption, and the joint likelihood of missing plus non-missing cases can be maximized using standard numerical methods (see Allison, 2002, p. 80).

[55] Some models can be ruled out through inspection of missing data patterns or by bounding the logical extremes of true impact compatible with the portion of the data that is observed.

However, this approach will produce biased results if any of the key assumptions behind the method do not hold. For example, if the selection process cannot be represented simply as "the data are observed if L <=0 and missing if L > 0," or if L and Y are not normally distributed, then Heckman's approach will yield biased impact estimates. Unfortunately, the approach is highly sensitive to the normality assumption (Little & Rubin, 2002; Stolzenberg & Relles, 1990, 1997; Allison, 2002). One version of the approach, known as the two-step estimator, is less sensitive to the normality assumption. However, it requires one additional assumption, known as an "exclusion restriction:" there must be one or more variables that influence the missing data rate that do not affect the value of the outcome, Y. This assumption is often problematic since in practice almost any reason one can think of for outcome data to be missing in educational RCTs is plausibly related to student achievement—the key outcome measure—and thus belongs as a covariate in the impact model. Other strategies for relaxing the normality assumption using semi-parametric methods have not fully removed this concern.[56]

Pattern-Mixture Models

An extension of maximum likelihood (ML) estimation─called pattern-mixture modeling─is often suggested to deal with NMAR cases where missingness of outcome data is not at random even conditional on background variables.

Pattern-mixture modeling postulates that every distinctive pattern of missing data on the Y, Trt, and X variables in a dataset represents a different *subpopulation* of the population. Each such subpopulation, S, is given its own joint p.d.f., f_S (Y, Trt, X), which is attached to the observations that exhibit that pattern. All the relationships among student outcomes, treatment assignment, and student/classroom/school/community background variables—that is, all the unknown parameters of f_S (Y, Trt, X)—are then allowed to take a different form for every different subpopulation/missing data pattern. To estimate the impact of an intervention, the key parameters appear in the first term of the usual decomposition of f_S (Y, Trt, X),

$$f_S(Y, Trt, X) = f_S(Y \mid Trt, X) \quad f_S(Trt, X),$$

since it is the conditional distribution of Y given Trt and X that provides information on the influence of the treatment assignment variable, Trt, on outcome Y.

Unfortunately, recognizing in this way that cases with missing data may differ systematically from other cases in terms of relationships among the variables rules out estimating many of the key parameters of interest without further assumptions (see Allison (2002), p. 82). In particular, it precludes estimation of the parameters of the overall conditional distribution of Y given Trt and X defined by the product of all of the separate f_S (Y │Trt, X) distributions. Many of these subpopulation-specific distributions cannot be estimated by ML methods, especially the ones that are defined by missingness on one or more of the three variables are not estimatable by ML methods.

For example, ML estimation will do a good job of estimating the parameters of f_S (Trt, X) for the subpopulation of students defined by complete data on Trt and X but missing data on Y, and make those parameters sensitive to what is different about those students—

[56] See for example Chamberlain (1986), Ahn, et al. (1993), Powell (1994), and Das, et al. (2003).

including differences on unobservable characteristics that condition relationships between Trt and the various variables in X (as well as among the individual X variables). But it can provide no information at all about the parameters of f_S (Y | Trt, X), which characterize the relationships between the Trt or X variables and Y. This is because *Y is never seen for any members of this subpopulation.* Thus, pattern mixture models replace the assumption of missingness at random conditional on observables with the more realistic assumption of non-random missing outcome data but only by introducing a set of non-estimatible parameters critical to obtaining the desired impact estimate once the "mixtures" in the pattern are put back together. (The literature calls these "non-identified" parameters.)

Allison (2002) provides a very simple yet highly illuminating example of the inherent problem. As he states, "To make *any* headway toward estimation, we must impose some restrictions on the [different] sets of parameters [for the different subpopulations]" (p. 82; emphasis added). He then imposes what Little (1993, 1994) calls "complete case missing-variable restrictions" as one way out of the quandary of non-identified parameters. Under this strategy, for the subpopulation defined by one variable but not all variables being missing, "the conditional distribution of the missing variable given the observed variable is equated to the corresponding [conditional] distribution for the complete-case pattern", by which he means the subpopulation for which there are *no* missing variables (p. 83). In other words, Little proposes to estimate the distributional parameters needed to calculate the intervention's impact for the subpopulation with complete data and then assume the same parameter values apply to other subpopulations that must play a part in producing one's overall impact finding but where nothing about those relationships can be observed. Transparently, this requires the assumption that missing data cases have no systematic differences from complete-data cases—i.e., that missing data are completely at random given the X and Trt variables.

Bounds and Sensitivity Tests

In our view, none of the methods described above effectively address unobserved characteristics that influence both the outcome Y and the probability of having missing data on Y. Hence, they are ineffective at addressing the NMAR case and—in the cases of selection modeling and pattern mixture modeling—may be misleading in appearing to have done so when they do not. Such methods rely on untestable assumptions and can be very sensitive in their findings to the particular assumptions made (Allison 2002). In most settings, there is no way to know how important unobserved factors are because we do not know why data are missing.

It is, however, vital for analysts to recognize that there are *two* separate sources of uncertainty in estimates of intervention impacts. The first is associated with selecting a random sample, where different samples may yield different estimates (i.e., sampling error), which is measured by the impact estimate's standard error. The second source of uncertainty is related to whether the data are really MAR or whether unobserved factors play a role—and, if so, how great a role. This second source of uncertainty can be large and is not reflected in the standard errors used to do statistical significance testing or construct confidence intervals. Ignoring it can lead to a false sense of security about the approach selected to "fix" the missing data problem, and too much confidence in the resulting intervention impact estimates.

In our view, a useful addition to impact estimation is the presentation of "bounds" around the impact estimates themselves to reflect the underlying uncertainty about the true missing data mechanism. This approach, due to Manski and Horowitz,[57] identifies *the range* of impact estimates consistent with the data if one is unwilling to make any assumptions about the mechanisms by which missing data are generated. For example, consider studies in which student proficiency is the outcome of interest. For studies that simply measure average outcomes rather than impacts, the bounding approach would measure average achievement under (1) the best-case scenario, assuming that *all* students with missing outcome data on achievement are proficient and (2) the worst-case scenario, assuming that *none* of the students with missing outcome data are proficient. This approach is purposely based on the most extreme situations that could occur in each direction to ensure that the true value of the parameter being estimated lies between (1) and (2).[58]

However, RCTs are not designed to measure average outcomes: they measure impacts. Consequently, in an RCT designed to measure the effects of, for example, a professional development program on student's math proficiency, we recommend computing: (1) a best-case impact estimate—i.e., the largest possible impact estimate consistent with the available data—by assuming that *all* treatment students with missing data are proficient and *no* control students with missing data are proficient; and (2) a worst-case impact estimate—i.e., the smallest possible impact estimate consistent with the available data—by assuming that *no* treatment students with missing data are proficient and *all* control students with missing data are proficient.[59]

The bounds implied by the best-case scenario and the worst-case scenario can be very wide if the rate of missing data is high. For example, suppose that math proficiency is missing for 15 percent of students in treatment schools and 20 percent of students in control schools. Exhibit 1 shows that in this example, the range between the best-case impact estimate and the worst-case impact estimate is 35 percentage points. One can similarly construct bounds for continuous outcome measures, such as scale scores or test scores of student achievement when these metrics have minimum and maximum values.

Furthermore, Exhibit 1 shows that case deletion or complete case analysis yields an impact estimate of +7.4 percentage points, while the worst-case impact estimate is -10 percentage points and the best-case estimate +25 percentage points. This illustrates the potential for uncertainty to emerge in a bounding analysis, and also points up the potential for the true impact to be very far from the estimate produced by the simplest missing data methodology, complete case analysis.

[57] See Manski (1990) and Horowitz & Manski (1998, 2000).

[58] More precisely, the width of the bounds reflects the amount of uncertainty due to the fact that the missing data mechanism is unknown. In principal, one could construct bounds that account both for sampling error and the uncertainty about the missing data mechanism. However, we have never seen this done in practice.

[59] More precisely, the "best-case impact estimate" and "worst-case impact estimate" provide the upper and lower bounds for true impact, if there were no sampling error.

Unfortunately, without additional information, there is no way to produce narrower bounds on the intervention's true impact. One can produce a point estimate of the treatment effect by making some standard assumption (e.g., that the data are Missing at Random). However, these types of assumptions are inherently untestable. Therefore, it is hard to be confident that any point estimate resulting from an untestable assumption is a better estimate of the treatment effect that any other point within the logical bounds.

Of course, impact estimates that fall *outside* the logical bounds must be incorrect. If a particular methodology produces a point estimate that falls outside of the logical bounds, its assumptions must be incorrect. Therefore, one useful feature of logical bounds is that they can be used to conduct "specification tests" of methods that make different assumptions.

An intermediate point between these two extremes—logical limits versus "pinpoint" assumptions—can help to make the bounding approach more useful by providing fewer possible impacts. This involves reducing, but not presuming to have eliminated, missing data uncertainty through further assumptions. One instance would assume that a share, s, of the missing values in the treatment group are missing completely at random (MCAR) or at random conditional on the covariates (MAR) and the remainder all equal "not proficient" on the outcome Y, and similarly for a share, r, of the missing values in the control group. One of the MAR imputation methods described above could then be used to generate unbiased imputed values for the s and r shares and the remaining missing cases addressed through Manski-Horowitz logical bounds (i.e., set to the extremes of "proficient" and "not proficient" in turn).[60] This would not put the lower bound on impact as low as the Manski-Horowitz bounds nor the upper bound as high, thereby tightening the policy inference drawn from the results. Though this is clearly better than assuming that *all* missing data are MCAR or MAR, this approach still adopts arbitrary assumptions for a share of the data.

[60] The authors are indebted to Jeffrey Smith for this suggestion.

Exhibit 1: Numeric Example of Bounding Approach to Estimating Impacts on Dichotomous Outcomes

A. Observed Data

Y_{obs} = Actual Math Proficiency

		0	1	missing
T = Treatment	0	20	60	20
	1	15	70	15

A complete case analysis yields an impact estimate of [70 / (70 + 15)] - [60 / (60 + 20)] = 0.074, or positive 7.4 percentage points.

B. Best-Case Scenario for the Treatment

Y_B = Math Proficiency Under Best-Case Scenario

		0	1	
T = Treatment	0	20 + 20 = 40	60	▨
	1	15	70 + 15 = 85	▨

The best-case scenario yields an impact estimate of [85 / (85 + 15)] - [60 / (60 + 40)] = 0.25, or +25 percentage points.

C. Worst-Case Scenario for the Treatment

Y_W = Math Proficiency Under Worst-Case Scenario

		0	1	
T = Treatment	0	20	60 + 20 = 80	▨
	1	15 + 15 = 30	70	▨

The worst-case scenario yields an impact estimate of [70 / (70 + 30)] - [80 / (80 + 20)] = -0.10, or -10 percentage points.

Alternatively, one could follow Altonji, et al. (2005) and posit that the observed covariates on which one is able to condition are a random subset of the full set of potential determinants of non-response (i.e., of the complete set of conditioning variables one would need to use to generate unbiased impact estimates).[61] This again seems arbitrary to us and not to be recommended.

Another approach introduced by DiNardo, et al. (2006) uses random variation in the missing data rate on Y to learn more about non-response bias. For example, using data from the Moving to Opportunities experiment,[62] wherein a subset of initial follow-up survey non-respondents were randomly selected and subjected to more intense data collection efforts, resulting in lower end-stage missing data rates on Y for those cases. By applying selection models, DiNardo and colleagues were able to use outcome information for members of the "swing group" whose Y values were observed only because of the enhanced data collection effort to extrapolate to all non-respondents and obtain a measure of impact on the entire treatment group. Unfortunately, as the authors acknowledge, this extrapolation works only when one makes assumptions about joint normality and/or "exclusion restrictions" similar—if less strong than—those required of selection modeling to deal with missing data absent variation in response rates induced by randomized follow-up.

DiNardo, et al. (2006) provides a more promising approach by extending the Manski-Horowitz bounding framework an added step. Here, the authors assume that assignment to treatment can only affect survey non-response in a single direction—either make the observation of Y more likely for all study participants or make it less likely for all participants. If this is the case, a potentially narrower set of bounds can be calculated for the impact of the intervention on the subset of individuals who will respond whether in the treatment group or the control group (the "complier" subpopulation). While this may be a useful range to report (along with the straight Manski-Horowitz bounds), it should be accompanied by the share of the entire study population of interest to which the range applies (a figure that equals the lower of the two response rates in the treatment and control groups). In addition, it would be prudent to hypothetically explore how treatment might increase or decrease response rates for some types of individuals, which would constitute a violation of the key assumption of the method.

Finally, it may be possible for education researchers to develop "consensus bounds"—bounds that rule out estimates that are logically possible, but which a consensus among content experts says cannot occur. What are needed here are bounds tighter than the logical bounds that reflect the range of impact estimates that the RCT could "reasonably" have generated. For example, suppose the study were using a vertically scaled test, and the outcome were a measure of test score gains. There may be a consensus that the lowest plausible value of the test score *gain* for students with missing post-tests is zero—that is, that negative gains are simply not plausible (and perhaps never observed in the

[61] Additional assumptions in the approach are that there are a great many such conditioning factors that contribute to the probability of non-response for specific observations, and that none of them dominates the non-response generating process as a whole.

[62] DiNardo, J, J. McCrary, & L. Sanbonmatsu (2006).

complete data). Effectively, this is equivalent to assuming that their follow-up test score is no lower than their pretest score. Where consensus can be built from widely credited information from outside the evaluation, it may be all the more convincing.

In the end, however, a wide range between the best- and worst-case impact estimates under the Manski-Horowitz bounding approach tells us that we cannot be totally confident about anything we say about the magnitude or even the direction of true impact. This is true unless we are confident that the data are missing completely at random or—if we adjust for observable background characteristics—missing at random—or at least close to it. We believe this is the right place to end up when providing guidance to policy makers and education practitioners who seek *evidence*, not speculation or assumptions or assertion, as the basis for their policy decisions. Logical bounds *provide decision-makers with all of the hard evidence that is available* from an RCT with missing outcome data; there is no way around this truism.

When complemented by consensus bounds and a point estimate that reflects the author's best attempt to address the bias that can result from missing data, the presentation of logical bounds seems to us the most honest and appropriate way to convey what one has learned from a given social experiment. Furthermore, the discipline of *not stating findings* derived through untestable, potentially strong assumptions when RCT data are missing communicates in the strongest possible terms the need for more complete data collection in future RCTs through wise study design and investments in high response rates, especially for outcome data and key covariates.

4. Testing the Performance of Selected Missing Data Methods

To provide further guidance to education researchers regarding what to do about missing data in an RCT, we conducted simulations to test the performance of selected missing data analysis methods described in Chapter 3. The simulations were run under conditions that varied on three dimensions: (1) the amount of missing data assumed, relatively low (5% missing) vs. relatively high (40% missing); (2) the level at which data are missing─at the level of schools (the assumed unit of randomization) and for students missing within schools; and, (3) the previously discussed underlying missing data mechanisms—i.e., MCAR, MAR, and NMAR. The benefit of conducting such simulations is that we know the true impact of our hypothetical intervention, and this allows us to compare the magnitude and precision of the estimated impacts produced by the different missing data methods under these varying conditions. This chapter begins with a description of the simulation methodology (greater details can be found in Appendices C and E), and then summarizes the simulation results (complete results are provided in Appendix D).

A. Simulation Methods

The simulations involved four steps: (1) developing parameters to define a hypothetical, but typical, educational RCT; (2) creating simulated data for the hypothetical RCT; (3) specifying different missing data mechanisms and test conditions; and, (4) implementing the selected missing data methods for both missing covariates and outcome measures. Each step is described below, followed by a summary of the results.

A Hypothetical Education RCT

To test the different missing data methods, we created a hypothetical, yet common, education RCT intended to measure the impacts of a particular intervention on student achievement in which schools are randomly assigned to two equal-size groups: (1) a treatment group, which receive some unspecified classroom or school-level intervention, or (2) a control group, which does not receive the intervention. Key features of this fictional RCT design include the following:

- a sample of 60 schools with 30 schools assigned to treatment and 30 schools assigned to control (i.e., the probability of assignment to treatment is 50 percent);
- 60 students in each school, for a total sample of 3,600 students;
- baseline data are available for students on gender, an unspecified risk status variable (e.g., high income versus low income), and pretest achievement data in a single subject area (either reading or mathematics); and,
- follow-up outcome data on achievement in the same subject area as the pretest.

For estimating the average impacts of the treatment on student achievement, we assumed a linear model of the post-test score on the treatment indicator and a set of control variables. All of the models included controls for gender and risk status, but we varied the analysis by including the pretest score in some models and not in others. The decision to collect pretest scores is an important design consideration in any evaluation, and the conventional wisdom suggests that the primary reason to collect pretest scores is to increase the precision of the impact estimates (e.g., Schochet, 2005). However, because the missing data mechanism may sometimes depend on the value of the pretest, and the pretest is such an important predictor of post-test scores, there is good reason to believe that controlling for pretest scores in the analysis may reduce bias from missing data. Therefore, in the simulations where some of the post-test scores are missing, we tested this hypothesis by estimating the analysis model with and without controlling for pretest scores.

To account for the nested structure of the data, we estimated standard, two-level hierarchical models of student achievement. The student-level model (Level 1) includes fixed effects for gender, risk status, and, in some simulations, the students' pretest score. The school-level model (Level 2) includes fixed effects for treatment assignment and random effects for individual schools (r_j in the equation below). More specifically, the following models were used in the simulations:

Level 1 Model (Students):

$$Y_{ij} = \beta_{0j} + \beta_1\left(SEX_{ij}\right) + \beta_2\left(RISK_{ij}\right) + \beta_3\left(PRETEST_{ij}\right) + e_{ij}$$

Level 2 Model (Schools):

$$\beta_{0j} = \gamma_0 + \gamma_1(Trt_j) + r_j$$

The parameter γ_1 indicates the intention-to-treat (ITT) average effect of the intervention on student achievement.

Creating Simulated Data

To implement the simulations we needed to create a data set that would be generated from our hypothetical study. This was done using the following parameters that are, of course, never known in an actual RCT:

- **The size of the treatment effect.** We assumed that the treatment has an average effect size of 0.20, or 20 percent of a standard deviation of the outcome for control group students. However, we also wanted the treatment effect to vary across students reflecting the goal of many education interventions to reduce achievement gaps. Therefore, the simulated data has an overall average positive treatment effect of 0.20, but the effect ranges from approximately zero for initially high achieving students (those with relatively higher pretest scores) to approximately 0.40 for initially low achieving students (those with relatively lower pretest scores).
- **The distribution of the control variables.** We constructed our two demographic control variables, gender and the unspecified risk factor, to be uncorrelated with each other. In addition, we assumed that both of these variables were correlated with the pretest measure. In particular, we set the parameters of the data generating process to ensure that:
    - **Average pretest scores are 0.20 standard deviations higher for girls than for boys.** For example, in the 2007 California Standards Test for

English/Language Arts, the average score for girls in the 3rd grade was 0.19 standard deviations higher than the average score for boys.[63] In addition, the National Assessment of Educational Progress shows a similarly sized gap between boys and girls at the national level.[64]

  - **Average pretest scores are 0.80 standard deviations higher for low-risk students than for high-risk students.** For example, in the 2007 California Standards Test for English/Language Arts, the average score for 3rd grade students eligible for free or reduced-price lunches was 0.79 standard deviations lower than the average score for other 3rd grade students.[65]

- **The relationship between control variables and the outcome variable.** We assumed that absent the intervention, the correlation between pretest and post-test scores was 0.50. In addition, we assumed that on average the pretest to post-test gain for girls was 0.02 standard deviation units more than boys, and that low-risk students gained 0.05 standard deviation units more than high-risk students.

- **The inter-class correlation.** In most education settings, students tend to be more similar to other students in the same school than to students in other schools. As a result, some of the variability in student achievement across students can be explained by variability across schools. For our simulations, we assumed an inter-class correlation (ICC) of 0.10 in pretest scores, which means that 10 percent of the variation in achievement across students can be explained by variation in mean pretest scores across schools.

Missing Data Mechanisms and Test Conditions

To test the selected methods, we took the simulated data and randomly made some of the data missing. In a real evaluation, we face missing data problems where the mechanism behind the missing data is unknown─only the amount of missing data and the association between missingness and the observed variables are revealed in the data. However, in our simulations, we varied several aspects of missing data so that we could compare the relative performance of the different methods under controlled conditions:

- **What data are missing?** In RCTs in education, the two key variables are the pretest and the post-test measures of achievement, and missing values in either of these variables may be especially problematic. Consequently, we conducted two sets of simulations—one in which the pretest score is missing for some fraction of the sample, and one in which the post-test score is missing for some fraction of the

[63] Average test scores for boys (324.4) and girls (335.7) can be found at the California Department of Education website (http://star.cde.ca.gov/star2007/Viewreport.asp). The pooled standard deviation in test scores for boys and girls (59) is provided in technical documentation (http://www.cde.ca.gov/ta/tg/sr/documents/csttechrpt07.pdf).

[64] Average test scores for boys (216) and girls (223) can be found using the NAEP Data Explorer (http://star.cde.ca.gov/star2007/Viewreport.asp). The pooled standard deviation in test scores for boys and girls (36) is also provided by the NAEP Data Explorer in technical documentation.

[65] Average test scores for students eligible for free or reduced price lunches (310.6) and other students (357.3) can be found at the California Department of Education website (http://star.cde.ca.gov/star2007/Viewreport.asp).

sample. For simplicity, we assumed that we have complete data on all of the other analytical variables.

- **How is missing related to treatment assignment?** For all of the simulations, we assumed that the rate of missing data was higher in the control group than in the treatment group (see, for example, Wolf, et al., 2009). This difference can often be explained by greater cooperation with data collection by treatment group schools and students compared to control schools and students.

- **What is the missing data mechanism?:** To test the performance of different missing data methods, we simulated data and tested different methods under the following three scenarios defined by the mechanism causing missing data:

  - **Scenario I: The missing data mechanism is entirely random within the treatment group and within the control group.** This is the most innocuous MCAR scenario.

  - **Scenario II: The missing data mechanism depends on treatment status and a student's pretest score.** More specifically, in this MAR scenario: (1) missingness was set to depend on both treatment and *pretest* scores; (2) missing data is more likely for students with lower pretest scores than for students with higher pretest scores, and (3) the relationship between missingness and pretest scores differs between the treatment and control groups.

  - **Scenario III: The missing data mechanism depends on treatment status and a student's post-test score.** More specifically, in this NMAR scenario: (1) missingness was set to depend on both treatment and *post*-test scores; (2) missing data is more likely for students with lower post–test scores than for students with higher post–test scores; and (3) the relationship between missingness and post-test scores differs between the treatment and control groups.

- **How much data are missing?:** The extent to which missing data can bias the estimated treatment effect may also depend on the amount of missing data. That is, if relatively few study participants have missing data, it may not matter which method is used to address the problem. However, if relatively many participants lack complete information, estimated impacts may be sensitive to the choice of analysis methods.

  Consequently, for each of these three scenarios, we simulated missing data at two relative extremes of missing data, a low of five percent missing outcome data and a high of 40 percent missing outcome data. These thresholds were set to extend beyond the experience of recent IES-sponsored RCTs in which missing outcome data has ranged from 10 to 20 percent.[66] In our experience, few evaluations have missing data rates of less than 5 percent for both the key outcome measures and key covariates, and few high-quality evaluations in education have missing data rates that are greater than 40 percent.

[66] See for example: Bernstein, et al., 2009; Campuzano, et al., 2009; Constantine, et al., 2009; Corrin, et al., 2008; Gamse, et al., 2009; Garet, et al., 2008; and, Wolf, et al., 2009.

In addition, for selected methods under one scenario, Scenario III, we also tested missing data rates *between* 5 percent and 40 percent to see how the performance of the methods varied with the missing data rate.

- **At what level are data missing?**: We ran a separate set of simulations for all the permutations described above but under the assumption that outcome data are missing for either 5 percent or 40 percent of *schools*—instead of for 5 percent or 40 percent of *students* within each school.

To better understand how we set values to missing, Exhibit 2 shows the missing data probabilities used to generate data with an overall missing data rate of 40 percent. It also shows how we set the probabilities higher for the control group than for the treatment group, and higher for students with lower achievement, as described earlier in this section.

Missing Data Methods Examined in the Simulations

For addressing missing data in the **pretest variable,** which is included as a control variable in some of our simulations, we tested the following methods described in Chapter 3:

- Case deletion (i.e., complete case analysis),
- Dummy variable adjustment,
- Mean value imputation,
- Non-stochastic regression imputation (including the post-test in the imputation model),
- Single stochastic regression imputation (including the post-test in the imputation model),
- Multiple stochastic regression imputation (including the post-test in the imputation model), and,
- Maximum likelihood─EM Algorithm with Multiple Imputaions.

For the dummy variable adjustment, we created a missing data dummy variable that equals one for the cases with missing pretest scores and zero otherwise, replaced missing pretest scores with zeros, and included the new dummy variable as a control variable in the model. For mean value imputation, we computed the mean value of the pretest for non-missing cases separately for the treatment group and the control group, and we replaced missing pretest scores with the respective group means.

Our approach to the three regression imputation methods varied depending on whether we were missing data on individual students or for entire schools. When data were **missing on individual students**, the imputation model included the student's gender, risk status, and post-test. In addition, we included school dummy variables to ensure that the imputed values captured the variability across schools. This model was estimated on non-missing cases and applied to missing cases to predict pretest scores, and we replaced the missing values with the predicted values from the model.

Exhibit 2: Missing Data Probabilities Under Three Different Scenarios: The Situation in Which the Overall Missing Data Rate = 40%

<table>
<tr><th colspan="3">Scenario I – Missingness Depends on Treatment Status Only (MCAR)</th></tr>
<tr><th>Quartile on Pretest Scores</th><th>Treatment Group</th><th>Control Group</th></tr>
<tr><td>1 (highest pretest scores)</td><td>35%</td><td>45%</td></tr>
<tr><td>2</td><td>35%</td><td>45%</td></tr>
<tr><td>3</td><td>35%</td><td>45%</td></tr>
<tr><td>4 (lowest pretest scores)</td><td>35%</td><td>45%</td></tr>
<tr><td><i>Average</i></td><td><i>35%</i></td><td><i>45%</i></td></tr>
<tr><th colspan="3">Scenario II - Missingness Depends on Treatment Status and Pretest Scores (MAR)</th></tr>
<tr><th>Quartile on Pretest Scores</th><th>Treatment Group</th><th>Control Group</th></tr>
<tr><td>1 (highest pretest scores)</td><td>30%</td><td>30%</td></tr>
<tr><td>2</td><td>35%</td><td>40%</td></tr>
<tr><td>3</td><td>35%</td><td>50%</td></tr>
<tr><td>4 (lowest pretest scores)</td><td>40%</td><td>60%</td></tr>
<tr><td><i>Average</i></td><td><i>35%</i></td><td><i>45%</i></td></tr>
<tr><th colspan="3">Scenario III - Missingness Depends on Treatment Status and Post-test Scores (NMAR)</th></tr>
<tr><th>Quartile on Post-test Scores</th><th>Treatment Group</th><th>Control Group</th></tr>
<tr><td>1 (highest post-test scores)</td><td>30%</td><td>30%</td></tr>
<tr><td>2</td><td>35%</td><td>40%</td></tr>
<tr><td>3</td><td>35%</td><td>50%</td></tr>
<tr><td>4 (lowest post-test scores)</td><td>40%</td><td>60%</td></tr>
<tr><td><i>Average</i></td><td><i>35%</i></td><td><i>45%</i></td></tr>
</table>

For stochastic regression imputation, we added a stochastic, student-level error term to the predicted value from the model by selecting randomly (and with replacement) from the residuals from the regression on non-missing cases.[67] For multiple imputation, we implemented stochastic regression imputation five times for each missing value,[68] estimated five different impact estimates—one for each imputation—and combined the results using Rubin's rules (e.g., Rubin, 1987, 1996).

As noted above, when data on individual students were missing, we included school dummy variables in the imputation models to take advantage of the fact that while we did not know pretest scores for X percent of students, we did know what schools they came from—and could estimate the school-effect from the other students for which pretest data were available. However, when pretest scores are **missing for entire schools**, this approach is not feasible. For example, when data are missing for 40 percent of schools, the data on the other 60 percent of schools cannot provide any information that is useful in estimating the school effects in the missing schools.

Therefore, when data were set to missing for entire schools, our imputation strategy involved calculating school-level means for each variable from the available data, imputing mean pretest scores for the schools with missing data, and estimating a one-level model from the school means to estimate the average treatment effect. Stochastic regression imputation from school-level means added school-level residuals to the predicted values from the imputation model, to ensure that the imputed values capture the between-school variability in outcomes that are present in the true values of the data. Furthermore, when schools are randomized instead of students, one can measure the treatment effect by estimating a school-level model of the school's mean outcome on a treatment indicator and school-level means of the individual-level covariates.

The EM algorithm with multiple imputation method was implemented in a manner very similar to that described for multiple stochastic regression imputation.[69] The difference being that in the latter case the imputed values were the predicted values from a regression model, and in the EM approach, the EM algorithm was used to obtain imputed values. In both approaches we generated five imputed data sets, and in both a random residual was added to each predicted value so that the imputed values in each of the five data sets would be slightly different from one another.

For addressing missing data in the **post-test variable,** we tested the following methods described in Chapter 3:

- Case deletion,
- Mean value imputation,

[67] Effectively, our imputation strategy treated the schools as fixed, while the analysis model treated the schools as random.

[68] The literature suggests that 5-10 imputations are adequate (see Rubin, 1987, 1996 and Little & Rubin, 2002).

[69] Of the different methods of maximum likelihood estimation available, multiple imputation using the EM algorithm was judged the most useful to test in the simulation analysis since it (a) can be implemented with all types of missing data (not just hierarchical missing data), (b) gives correct standard errors for impact estimates, (c) allows estimation of the two-level random intercepts impact model frequently used in educational RCTs, and (d) does not require specialized computer software or expertise. See Chapter 3 for more details.

- Non-stochastic regression imputation (including the pretest in the imputation model),
- Single stochastic regression imputation (including the pretest in the imputation model),
- Multiple stochastic regression imputation (including the pretest in the imputation model),
- Simple weighting approach using the inverse of observed response rates,
- More sophisticated weighting approach that involved modeling non-response to create weights,
- Fully-specified regression models with treatment/covariate interactions, and,
- Maximum likelihood—EM algorithm with multiple imputation.

The regression imputation methods, and the maximum likelihood method, were implemented in the same way as described above for missing pretests with one exception: we included the pretest in the model to impute the post-test. The simple weighting approach, designed to ensure that schools with high missing data rates are not underrepresented in the analysis, involves weighting each student with non-missing data by the inverse of the response rate in the same school.[70] The more sophisticated method involved the following steps: (1) estimate a logit model of data availability (1=non-missing post-test and 0=missing post-test) as a function of all of the available covariates, (2) divide the sample into quintiles based on the estimated probabilities of non-missing data, (3) compute response rates for each quintile, and (4) weight each student in a quintile by the inverse of the response rate for that quintile. For the fully-specified regression models with treatment/covariate interactions, we included the interaction between treatment and the pretest score in the analysis model, and we evaluated the impact at the mean of the pretest.[71]

B. Simulation Results

This section begins with a discussion of the standards we used to assess the relative performance of the different missing data methods that we tested, and then provides a summary of the simulation results (complete results are provided in Appendix D).

Assessing the Performance of the Different Methods

To judge the performance of the different missing data methods, we assessed the extent to which each approach produced bias in the impact estimate that would be considered "high" relative to the benchmark set by the *What Works Clearinghouse* (WWC). In the WWC, RCTs with attrition rates that are likely to yield non-response bias of 0.05 standard deviations or greater are treated as if they were quasi-experimental studies and are required to provide additional evidence suggesting that impact estimates are unbiased

[70] Therefore, if we had post-test scores for 40 of 60 students in a particular school, each of the 40 "respondents" would receive a weight of $(40/60)^{-1}$ or 1.5.

[71] We did not interact the treatment indicator with other covariates. Because the simulated missing data mechanism is a function of the pretest—but unrelated to the other covariates after conditioning on the pretest—there is no reason to expect that the estimates would be any different if we had included interactions with all of the covariates in the model.

(U.S. Department of Education, 2008). Because the RCTs in education that are currently underway may at some point be subject to review by the WWC, we decided to accept the 0.05 standard deviation threshold for bias in assessing the performance of different missing data methods. In each of our simulations, methods that yielded bias in the impact estimate of greater than 0.05 standard deviations were deemed to have produced "high bias," while methods that yielded bias of less than 0.05 standard deviations were deemed to have produced "low bias."

Additionally, some of the methods may also yield biased standard errors which contribute to the hypothesis test of whether the impact estimate is statistically significant.[72] Therefore, we decided it was also important to set standards for assessing the magnitude of the bias in the estimates of the standard errors. We classified the bias in a standard error estimate as large ("high bias") if it would generate as much bias in the t-statistic as is produced a 0.05 standard deviation bias in the impact estimate itself. In this way, we rely entirely on the WWC's attrition standard to determine whether the bias in the impact estimate or standard error should be treated as large ("high bias") or small ("low bias"). For more details on how we calculated the bias thresholds for the standard errors, see Appendix E.

Simulation Results

Exhibit 3 summarizes the results from the simulations in which data were missing from 40 percent of students within each school; Exhibit 4 summarizes the results from the simulations in which data were missing from 40 percent of schools. Each table presents the two key performance measures: (1) bias in the impact estimate, and (2) bias in the estimated standard error. The tables include three columns, one for each of the three scenarios—Scenario I, in which the data were missing at random within group (treatment or control); Scenario II, in which the data were missing at random after conditioning on group *and* pre-intervention characteristics of the students (demographics and pretest scores); and Scenario III, in which the missing data depended on the outcome measure—student post-test scores—even after conditioning on group and pre-intervention characteristics of the students.

As discussed above, we also conducted simulations in which data were missing for five percent of students and schools, but none of the methods produced bias that exceeded the thresholds that we selected for these simulations under any of the three scenarios and for either missing pretests or post-tests. Therefore, we do not provide summary tables for the results from these simulations (the results themselves are provided in Appendix D).

[72] The t-statistic equals the estimated impact divided by the estimated standard error.

Exhibit 3: Summary of Simulation Results, Missing Data for 40% of Students

		Impact Estimate				Standard Error of Impact Est.				Overall
Data	**Pretest Data Available?**	**Scenario I**	**Scenario II**	**Scenario III**	**Low Bias in All Three Scenarios?**	**Scenario I**	**Scenario II**	**Scenario III**	**Low Bias in All Three Scenarios?**	**Low Bias for Both Estimates in All Scenarios?**
No Missing Data	No	Low Bias	Low Bias	Low Bias	√	Low Bias	Low Bias	Low Bias	√	√
	Yes	Low Bias	Low Bias	Low Bias	√	Low Bias	Low Bias	Low Bias	√	√
A. Pretest (X) Data Missing										
Case Deletion	Yes	Low Bias	Low Bias	High Bias		Low Bias	Low Bias	Low Bias	√	
Dummy Variable Method	Yes	Low Bias	Low Bias	Low Bias	√	Low Bias	Low Bias	Low Bias	√	√
Mean Value Imputation	Yes	Low Bias	High Bias	High Bias		Low Bias	Low Bias	Low Bias	√	
Single, Non-stochastic RI	Yes	Low Bias	High Bias	Low Bias		Low Bias	Low Bias	Low Bias	√	
Single, Stochastic RI	Yes	Low Bias	Low Bias	Low Bias	√	Low Bias	Low Bias	Low Bias	√	√
Multiple, Stochastic RI (n = 5)	Yes	Low Bias	Low Bias	Low Bias	√	Low Bias	Low Bias	Low Bias	√	√
EM Algorithm with MI (n = 5)	Yes	Low Bias	Low Bias	Low Bias	√	Low Bias	Low Bias	Low Bias	√	√
B. Post-Test (Y) Data Missing										
Case Deletion	No	Low Bias	High Bias	High Bias		Low Bias	Low Bias	Low Bias	√	
	Yes	Low Bias	Low Bias	High Bias		Low Bias	Low Bias	Low Bias	√	
Mean Value Imputation	No	Low Bias	High Bias	High Bias		High Bias	High Bias	High Bias		
		Low Bias	High Bias	High Bias		High Bias	High Bias	High Bias		
Single, Non-stochastic	No	Low Bias	High Bias	High Bias		Low Bias	Low Bias	Low Bias	√	

Exhibit 3: Summary of Simulation Results, Missing Data for 40% of Students

		Impact Estimate				Standard Error of Impact Est.				Overall
Data	Pretest Data Available?	Scenario I	Scenario II	Scenario III	Low Bias in All Three Scenarios?	Scenario I	Scenario II	Scenario III	Low Bias in All Three Scenarios?	Low Bias for Both Estimates in All Scenarios?
Regression Imputation	Yes	Low Bias	Low Bias	High Bias		Low Bias	Low Bias	Low Bias	√	
Single, Stochastic	No	Low Bias	High Bias	High Bias		Low Bias	Low Bias	Low Bias	√	
Regression Imputation	Yes	Low Bias	Low Bias	High Bias		Low Bias	Low Bias	Low Bias	√	
Multiple, Stochastic	No	Low Bias	High Bias	High Bias		Low Bias	Low Bias	Low Bias	√	
Regression Imputation (n = 5)	Yes	Low Bias	Low Bias	High Bias		Low Bias	Low Bias	Low Bias	√	
EM Algorithm with Multiple Imputation	No	Low Bias	High Bias	High Bias		Low Bias	Low Bias	Low Bias	√	
(n = 5)	Yes	Low Bias	Low Bias	High Bias		Low Bias	Low Bias	Low Bias	√	
Weighting - Simple	No	Low Bias	High Bias	High Bias		Low Bias	Low Bias	Low Bias	√	
	Yes	Low Bias	Low Bias	High Bias		Low Bias	Low Bias	Low Bias	√	
Weighting - Sophisticated	No		Not Estimated				Not Estimated			
	Yes	Low Bias	Low Bias	High Bias		Low Bias	Low Bias	Low Bias	√	
Fully Specified Regression Models	No		Not Applicable				Not Applicable			
w/ Treatment-Covariate Interactions	Yes	Low Bias	Low Bias	High Bias		Low Bias	Low Bias	Low Bias	√	

Notes:
For more details on the simulations, see Chapter 4 and Appendix D.

Exhibit 4: Summary of Simulation Results, Missing Data for 40% of Schools

		Impact Estimate				Standard Error of Impact Est.				Overall
Data	Pretest Data Available?	Scenario I	Scenario II	Scenario III	Low Bias in All Three Scenarios?	Scenario I	Scenario II	Scenario III	Low Bias in All Three Scenarios?	Low Bias for Both Estimates in All Scenarios?
No Missing Data	No	Low Bias	Low Bias	Low Bias	√	Low Bias	Low Bias	Low Bias	√	√
	Yes	Low Bias	Low Bias	Low Bias	√	Low Bias	Low Bias	Low Bias	√	√
A. Pretest (X) Data Missing										
Case Deletion	Yes	Low Bias	Low Bias	Low Bias	√	Low Bias	Low Bias	Low Bias	√	√
Dummy Variable Method	Yes	Low Bias	Low Bias	Low Bias	√	Low Bias	Low Bias	Low Bias	√	√
Mean Value Imputation	Yes	Low Bias	Low Bias	Low Bias	√	Low Bias	Low Bias	Low Bias	√	√
Single, Non-stochastic RI	Yes	Low Bias	Low Bias	Low Bias	√	High Bias	High Bias	High Bias		
Single, Stochastic RI	Yes	Low Bias	Low Bias	Low Bias	√	High Bias	High Bias	High Bias		
Multiple, Stochastic RI (n = 5)	Yes	Low Bias	Low Bias	Low Bias	√	Low Bias	Low Bias	Low Bias	√	√
EM Algorithm with MI (n = 5)	Yes	Low Bias	Low Bias	Low Bias	√	Low Bias	Low Bias	Low Bias	√	√
B. Post-Test (Y) Data Missing										
Case Deletion	No	Low Bias	Low Bias	Low Bias	√	Low Bias	Low Bias	Low Bias	√	√
	Yes	Low Bias	Low Bias	Low Bias	√	Low Bias	Low Bias	Low Bias	√	√
Mean Value Imputation	No	Low Bias	Low Bias	Low Bias	√	High Bias	High Bias	High Bias		
	Yes	Low Bias	Low Bias	Low Bias	√	High Bias	High Bias	High Bias		

Exhibit 4: Summary of Simulation Results, Missing Data for 40% of Schools

Data	Pretest Data Available?	Impact Estimate			Low Bias in All Three Scenarios?	Standard Error of Impact Est.			Low Bias in All Three Scenarios?	Overall
		Scenario I	Scenario II	Scenario III		Scenario I	Scenario II	Scenario III		Low Bias for Both Estimates in All Scenarios?
Single, Non-stochastic	No	Low Bias	Low Bias	Low Bias	√	High Bias	High Bias	High Bias		
Regression Imputation	Yes	Low Bias	Low Bias	Low Bias	√	High Bias	High Bias	High Bias		
Single, Stochastic	No	Low Bias	Low Bias	High Bias		High Bias	High Bias	High Bias		
Regression Imputation	Yes	Low Bias	Low Bias	Low Bias	√	High Bias	High Bias	High Bias		
Multiple, Stochastic	No	Low Bias	Low Bias	Low Bias	√	Low Bias	Low Bias	Low Bias	√	√
Regression Imputation (n = 5)	Yes	Low Bias	Low Bias	Low Bias	√	Low Bias	Low Bias	Low Bias	√	√
EM Algorithm with Multiple Imputation	No		Not Estimated		√		Not Estimated		√	
(n = 5)	Yes	Low Bias	Low Bias	Low Bias	√	Low Bias	Low Bias	Low Bias	√	√
Weighting - Simple	No		Not Applicable				Not Applicable			
	Yes		Not Applicable				Not Applicable			
Weighting - Sophisticated	No		Not Estimated				Not Estimated			
	Yes	Low Bias	Low Bias	Low Bias	√	Low Bias	Low Bias	Low Bias	√	√
Fully Specified Regression Models	No		Not Applicable				Not Applicable			
w/ Treatment-Covariate Interactions	Yes	Low Bias	Low Bias	Low Bias	√	Low Bias	Low Bias	Low Bias	√	√

Notes:

For more details on the simulations, see Chapter 4 and Appendix D.

We used the simulation results to assess the performance of different missing data methods when applied to the specific context that our simulations were designed to inform—Group Randomized Trials in which schools are randomized to treatment or control. Below, we present the results for the different missing data methods in those simulations for which pretest scores were collected and included in the impact analysis model.[73]

Case Deletion. Although often criticized, the technical literature provides a more nuanced view of case deletion. For example, Allison (2002) indicates that case deletion will work well in some situations and poorly in others. More specifically, he indicates that case deletion will yield biased impact estimates when missing data for an independent variable depends on the observed value of the dependent variable (see Allison 2002, p. 6).[74] In our simulations, this scenario corresponds to missing pretest where the missingness depends on the post-test (Scenario III). However, he also indicates that case deletion yields less bias in the coefficient estimates than other methods when missing data on an independent variable depends on its unobserved value.[75] In our simulations, this corresponds to missing pretest where the missingness depends on the pretest (Scenario II).

The results from our simulations are consistent with Allison's assessment. When **pretest data were missing** for 40 percent of students, and missingness depended on the value of the post-test (Scenario III), case deletion yielded impact estimates with bias that exceeded 0.05 standard deviations. In contrast, most other methods yielded impact estimates with bias of less than 0.05 standard deviations (see Exhibits 3 and 4). In addition, under Scenario III, when pretest scores were missing for either 40 percent of students or 40 percent of schools, case deletion produced impact estimates with greater bias than all of the other methods we tested, except for mean value imputation (see Appendix D, Tables III.b.1 and III.b.2).

However, also consistent with Allison's assessment, when missing pretest scores depended on the value of the pretest itself (Scenario II), case deletion yielded impact estimates with bias of less than 0.05 standard deviations (see Exhibits 3 and 4). In this scenario, case deletion of missing students or schools produced impact estimates that were closer to the true impact of 0.20 than all of the other methods we tested (see Appendix D, Tables II.b.1 and II.b.2). Therefore, in summary, the simulation results for case deletion closely matched the results that the literature would lead us to expect for missing pretest scores, i.e., case deletion produced impact estimates with less bias than other methods under some conditions and more bias than other methods under other conditions.

For **missing post-test scores**, however, case deletion worked as well as, or better than, all of the alternative methods across all of the missing data scenarios. In most of the missing post test scenarios, this method produced impact estimates that were less biased than the thresholds set for the simulations, and in all scenarios the biases in standard errors were less than the WWC-based thresholds. In the simulations where this method produced impact estimates with bias that

[73] In addition, for simulations involving missing outcomes, pretest scores are used in the imputation models, in constructing weights, and—in testing fully specified regression models with treatment-covariate interactions—in constructing the interaction terms.

[74] Allison (2002) refers to the missing data mechanism in this situation as MAR because missingness in the independent variable depends only on observed data on the dependent variable.

[75] Allison (2002) refers to the missing data mechanism in this situation as NMAR because missingness in the independent variable depends on the latent or unobserved values of the independent variable.

exceeded 0.05 standard deviations (Scenario III, missing post-test scores for 40 percent of students), none of the other methods produced impact estimates with bias of less than 0.05 standard deviations.

Finally, with respect to bias in the impact estimates from missing post-test scores, case deletion performed similarly to other methods in the following sense: it only produced bias of greater than 0.05 standard deviations under Scenario III when data were missing for 40 percent of schools (see Exhibit 3). In all other simulations, it produced bias of less than 0.05 standard deviations (see Exhibits 3 and 4). In addition, for missing post-test scores, the difference in bias between the case deletion and other tested methods were, in all but one case, less than 0.01 (see Appendix D). For example, under Scenario II, when data were missing for 40 percent of students, multiple stochastic regression imputation yielded an impact estimate that exactly equals the true impact of 0.20, while case deletion yields an impact estimate that equals 0.193—a difference of 0.007 standard deviations.

In summary, case deletion produced bias in the impact estimates that exceeded the WWC-based threshold in two of our simulations:

- **When pretest scores were missing for 40 percent of students under Scenario III.** Under this scenario, case deletion produced impact estimates with bias of greater than 0.05 standard deviations, while most other methods produced impact estimates with bias of less than 0.05 standard deviations.
- **When post-test scores were missing for 40 percent of students under Scenario III.** Under this scenario, none of the methods produced impact estimates with bias of less than 0.05 standard deviations.

Dummy Variable Method. The dummy variable method has been criticized in the literature for producing biased coefficient estimates (Allison, 2002 and Jones, 1996). While Jones (1996) is commonly cited as evidence that this method yields biased estimates, the appendix to this journal article provides a proof that the coefficient estimates will be unbiased if the two independent variables in his example, the one with missing data and the one without missing data, are uncorrelated with each other. In RCTs the variable of interest is the treatment indicator, which is never missing. Furthermore, when data are complete, randomization ensures that the treatment indicator is uncorrelated with the other independent variables. This raises the question of whether the standard critique of the dummy variable method applies in the particular context of education RCTs—in particular, in Group Randomized Trials where schools are randomly assigned to treatment or control, but the pretest score (or some other covariate) is missing for some students or schools.

The evidence from the simulation results indicates that for **missing pretest scores**, the dummy variable method performed similarly to the more sophisticated methods. In particular, we found that the dummy variable method produced impact estimates with bias of less than 0.05 standard deviations under all three scenarios (see Exhibits 3 and 4). In addition, in none of our simulations did the dummy variable produce standard errors with bias that exceeded the threshold established for these simulations (see Appendix D). Therefore, our simulation results cast doubt on whether the general concerns about this method, which we do not dispute, should deter analysts from adopting it in studies that randomly assign schools to educational interventions.

In summary, the dummy variable method produced impact estimates and standard error estimates with bias that fell within the acceptable range, as defined by the WWC-based criteria that we selected, in all of our simulations.

Mean Value Imputation. In general, mean value imputation is known to produce biased estimates of the standard errors of coefficients in regression models (see Allison, 2002 and Haitovsky, 1968). While there is no particular reason to believe this conclusion would not apply to RCTs in which schools are randomly assigned to treatment or control, our simulations shed light on whether this method, when applied to missing pretest scores or missing post-test scores, yields standard error estimates (1) with more or less bias than other methods, and (2) with bias that exceeds the threshold we developed for these simulations.

When data were missing for **pretest scores**, mean value imputation did not produce standard error estimates with bias that exceeded the WWC-based thresholds chosen for these simulations. When data were missing for 40 percent of students, however, mean value imputation produced impact estimates with bias that exceeded 0.05 for Scenarios II and III (see Exhibits 3 and 4).

When data were missing for **post-test scores**, mean value imputation produced standard error estimates with bias that exceeded the WWC-based thresholds in many of our simulations (see Exhibits 3 and 4). In fact, mean value imputation was the only method to yield standard errors with bias that exceeded the chosen thresholds in all three scenarios when data were missing for 40 percent of students and when data were missing for 40 percent of schools (see Exhibits 1 and 2). Finally, it is worth noting that when data were missing for 40 percent of students, mean value imputation was the only method to yield bias of greater than 0.05 standard deviations under both Scenarios II and III.

In summary, mean value imputation produced bias in the impact estimates and standard errors that exceeded the WWC-based thresholds in several of our simulations:

- **When pretest scores were missing for 40 percent of students under Scenario II.** Under this scenario, mean value imputation was one of two methods to produce impact estimates with bias greater than 0.05 standard deviations.
- **When pretest scores were missing for 40 percent of students under Scenario III.** Under this scenario, mean value imputation produced impact estimates with bias of greater than 0.05 standard deviations, while most of the other methods we tested produced impact estimates with bias of less than 0.05 standard deviations.
- **When pretest scores were missing for 40 percent of schools.** Under all three scenarios, when pretest scores were missing for 40 percent of schools, mean value imputation produced standard error estimates with bias that exceeded the WWC-based threshold.
- **When post-test scores were missing for 40 percent of students under Scenario II.** Under this scenario, mean value imputation was the only method to produce impact estimates with bias of greater than 0.05 standard deviations.
- **When post-test scores were missing for 40 percent of students under Scenario III.** Under this scenario, none of the methods produced impact estimates with bias of less than 0.05 standard deviations.

Single Non-Stochastic Regression Imputation. In general, single non-stochastic regression imputation is well-known to yield standard error estimates that are biased downward (see

Chapter 3). When used to impute missing pretest or post-test scores in a Group Randomized Trial, our simulation results can be used to address the question of whether this method yields standard error estimates (1) with more or less bias than other methods, and (2) with bias that exceeds the threshold chosen for these simulations.

When either **pretest or post-test scores were missing for 40 percent of schools**—the unit of random assignment—single non-stochastic regression imputation produced standard error estimates with bias that exceeded the WWC-based threshold (see Exhibit 4). In fact, when pretest scores were missing for 40 percent of students, the estimated bias was greater for this method than for any of the other methods.

However, when either **pretest or post-scores were missing for 40 percent of students** within each school, single non-stochastic regression imputation produced standard error estimates with bias that fell below the WWC-based threshold (see Exhibit 3). In fact, the estimated bias was less than or equal to 0.001 standard deviations, or one percent of the true standard error, in all three simulations. This suggests that when schools are randomly assigned but data are missing at the student level, the general concerns about single non-stochastic regression imputation may not apply.

In summary, single non-stochastic regression imputation produced bias in the impact estimates and standard errors that exceeded the WWC-based thresholds in several of our simulations:

- **When pretest scores were missing for 40 percent of students under Scenario II.** Under this scenario, single non-stochastic regression imputation was one of two methods to produce impact estimates with bias of greater than 0.05 standard deviations.
- **When pretest scores were missing for 40 percent of schools.** Under all three scenarios, when pretest scores were missing for 40 percent of schools, single non-stochastic regression imputation produced standard error estimates with bias that exceeded the WWC-based threshold.
- **When post-test scores were missing for 40 percent of students under Scenario III.** Under this scenario, none of the methods produced impact estimates with bias of less than 0.05 standard deviations.
- **When post-test scores were missing for 40 percent of schools under all three scenarios.** Under all three scenarios, when post-test scores were missing for 40 percent of schools, single non-stochastic regression imputation produced standard error estimates with bias that exceeded the WWC-based threshold.

Single Stochastic Regression Imputation. Single stochastic regression imputation is considered to be a "partial fix" to the problem associated with single non-stochastic regression imputation (see Chapter 3). Therefore, we would expect the bias in the standard error to be lower with single stochastic regression imputation than with single non-stochastic regression imputation.

Our simulation results indicate that *relative to the WWC-based threshold for bias in the standard error*, single stochastic regression imputation performed equally to single non-stochastic regression imputation. By this, we specifically mean that in each simulation (e.g., for both missing pretests and missing post-tests in all three scenarios), both methods produced standard errors that either exceeded the bias threshold or fell below the threshold (see Exhibits 3 and 4).

However, these results should *not* be interpreted as evidence against the conclusion from the literature that single stochastic regression imputation produces standard errors with less bias than single non-stochastic regression imputation. For each of the simulations with missing data for 40 percent of *schools*, the estimated bias was smaller for single stochastic regression imputation than for single non-stochastic regression imputation (see Appendix D, Tables I.b.2, II.b.2, and III.b.2).

Finally, with respect to bias in the impact estimates themselves, and relative to the WWC-based threshold, Exhibits 3 and 4 show that single stochastic regression imputation performed equivalently to most other methods, including single non-stochastic regression imputation. In all but one of the simulations, the bias in the impact estimate was either greater than 0.05 standard deviations for both of these two methods or less than 0.05 standard deviations for both methods.[76] This is not surprising since the addition of a stochastic error term to the imputed values is not intended to reduce bias in the impact estimate; rather, it is intended to reduce bias in the estimated standard error of the impact estimate.

In summary, single stochastic regression imputation produced bias in the impact estimates and standard errors that exceeded the WWC-based thresholds in some of our simulations:

- **When pretest scores were missing for 40 percent of schools.** Under all three scenarios, when pretest scores were missing for 40 percent of schools, single stochastic regression imputation produced standard error estimates with bias that exceeded the WWC-based threshold.
- **When post-test scores were missing for 40 percent of students under Scenario III.** Under this scenario, none of the methods produced impact estimates with bias of less than 0.05 standard deviations.
- **When post-tests score were missing for 40 percent of schools under all three scenarios.** Under all three scenarios, when post-test scores were missing for 40 percent of schools, single stochastic regression imputation produced standard error estimates with bias that exceeded the WWC-based threshold.

Multiple Stochastic Regression Imputation. Multiple stochastic regression imputation is considered to be a technically appropriate solution to the problem associated with single stochastic regression imputation (see Chapter 3). Therefore, we would expect the bias in the standard error to be either low or zero—and lower than the bias from single stochastic regression imputation.

The simulation results were consistent with this expectation. In all of our simulations, multiple stochastic regression imputation produced standard errors with bias estimates that fell below the WWC-based threshold selected for these simulations (see Exhibits 3 and 4), including the scenarios where both of the single regression imputation methods produced bias that exceeded the WWC-based threshold (see Exhibit 4).

With respect to bias in the impact estimates themselves, and relative to the WWC-based threshold, Exhibits 3 and 4 show that multiple stochastic regression imputation performed

[76] The exception was Scenario III with missing post-test scores for 40 percent of schools, where the bias was greater than 0.05 standard deviations for single stochastic regression imputation but less than 0.05 for single non-stochastic regression imputation. However, there is no reason to expect systematic differences between these two methods, so this difference can be attributed to random chance.

equivalently to most other methods, including single stochastic regression imputation. In all but one the scenarios, the bias in the impact estimate was either greater than 0.05 standard deviations for both of these two methods or less than 0.05 standard deviations for both methods.[77] This is not surprising since multiple imputation is not designed to produce less biased impact estimates than single stochastic regression imputation: it is intended to reduce bias in the estimated standard error of the impact estimate.

In summary, multiple stochastic regression imputation produced bias in the impact estimates that exceeded the WWC-based thresholds in only one of our simulations: when post-tests scores were missing for 40 percent of students under Scenario III. Under this scenario, none of the methods produced impact estimates with bias of less than 0.05 standard deviations.

EM Algorithm with Multiple Imputation. As discussed in Chapter 3, the EM algorithm is a maximum likelihood approach that can be used to directly obtain coefficient estimates or to impute missing values. When combined with multiple imputation, the literature suggests this approach should yield standard errors with little or no bias (see Chapter 3).

The simulation results are consistent with this expectation. Our simulation results indicate that *relative to the WWC-based threshold for bias in the standard error*, the EM algorithm with multiple imputation performed equally to multiple stochastic regression imputation. When data were missing for 40 percent of students or schools, the EM algorithm with multiple imputation produced standard errors with estimated bias that fell below the WWC-based threshold selected for these simulations in all three scenarios and for both missing pretests and missing post-tests (see Exhibits 3 and 4).

In addition, with respect to bias in the impact estimates themselves, relative to the WWC-based threshold, Exhibits 3 and 4 show that the EM algorithm with multiple imputation performed equivalently to most other methods, including multiple stochastic regression imputation. Like multiple stochastic regression imputation, when data were missing for 40 percent of students, the EM algorithm with multiple imputation produced impact estimates with bias of less than 0.05 when the missing data mechanism could be characterized as MCAR or MAR (e.g., Scenarios I and II for missing post-test scores), and it produced impact estimates with bias of greater than 0.05 when the missing data mechanism could be characterized as NMAR (e.g., Scenario III for missing post-test scores). When data were missing for 40 percent of schools, the EM algorithm with multiple imputation produced impact estimates with bias of less than 0.05 standard deviations for both missing pretests and post-tests under all three scenarios.

In summary, the EM algorithm with multiple imputation produced bias in the impact estimates that exceeded the WWC-based thresholds in only one of our simulations: when post-tests scores were missing for 40 percent of students under Scenario III.

Weighting, Simple Approach. The simple weighting approach, which can be applied in evaluations in which data are missing for selected students in each school, involves weighting up the students with non-missing data to the count of all students in the school. If the impact of the intervention varies across schools, this method might be expected to produce impact estimates with less bias than case deletion. If impact of the intervention does not vary across schools, then we might expect this method to produce impact estimates with bias that is equivalent to that of case deletion. In our simulation scenarios, the true impact was constant across schools.

[77] Ibid., footnote 77.

Therefore, we would not expect this method to produce impacts that are much different from the impacts produced by case deletion.

The simulation results are consistent with these expectations. Relative to the bias standards that we adopted for both impacts and standard errors, the performance of the simple weighting approach was equivalent to the performance of case deletion for all simulations (see Exhibits 3 and 4). In addition, the difference in impacts between the simple weighting approach and case deletion was less than or equal to 0.003 standard deviations in all of the simulations (see Appendix D, Tables I.b1, II.b.1, and III.b.1).

In summary, the simple weighting approach produced bias in the impact estimates that exceeded the WWC-based thresholds in only one of our simulations: when post-tests scores were missing for 40 percent of students under Scenario III.

Weighting, More Sophisticated Approach. As described earlier, this approach involves estimating a propensity model and using this model to assign weights to cases with non-missing data. In the literature, this method is considered an acceptable alternative to multiple imputation.

Relative to the bias standards that we adopted for both impacts and standard errors, the performance of the more sophisticated weighting approach was equivalent to the performance of both multiple stochastic regression imputation and the EM algorithm with multiple imputation for all simulations (see Exhibits 3 and 4). In addition, the difference in impacts between the more sophisticated weighting approach and multiple stochastic regression imputation was less than or equal to 0.001 standard deviations in simulations with missing data for 40 percent of students (see Appendix D, Tables I.b1, II.b.1, and III.b.1) and less than or equal to 0.01 standard deviations in simulations with missing data for 40 percent of schools (see Appendix D, Tables I.b.2, II.b.2, and III.b.2).

In summary, the more sophisticated weighting approach produced bias in the impact estimates that exceeded the WWC-based thresholds in only one of our simulations: when post-tests scores were missing for 40 percent of students under Scenario III.

Fully Interacted Regression Models with Treatment-Covariate Interactions. As described earlier in this chapter, we tested this approach by adding the interaction between the treatment indicator and the pretest variable as an independent variable in the model used to estimate the impacts of the intervention. This method ensures that the average treatment effect is evaluated at the mean for the entire sample—not just the mean for the sample with complete post-test data. Because of this, we would expect this method to produce impact estimates with less bias than case deletion. However, we had no prior expectations regarding the expected performance of this method relative to the other methods.

Relative to the bias standards that we adopted for both impacts and standard errors, the performance of this method was equivalent to the performance of the methods we have recommended thus far (see Exhibits 3 and 4). In addition, the difference in impacts between fully interacted regression models with treatment-covariate interactions and multiple stochastic regression imputation was less than or equal to 0.002 standard deviations in simulations with missing data for 40 percent of students (see Appendix D, Tables I.b1, II.b.1, and III.b.1) and less than or equal to 0.009 standard deviations in simulations with missing data for 40 percent of schools (see Appendix D, Tables I.b.2, II.b.2, and III.b.2).

In summary, fully interacted regression models with treatment-covariate interactions produced bias in the impact estimates that exceeded the WWC-based thresholds in only one of our simulations: when post-tests scores were missing for 40 percent of students under Scenario III.

Testing a Range of Missing Data Rates

As discussed above, our simulations tested the performance of selected missing data methods at two levels of missing data for schools and students, i.e., we ran the simulations at five percent and 40 percent missing, respectively. This raised an obvious question, "Is there a point along this range of possible attrition at which the results change?" To explore the sensitivity of the results to intermediate missing data rates, we ran simulations within the 5%-40% range for a subset of missing data methodologies. In particular, for missing post-test data and Scenario III—the scenario that analysts worry the most about because the data are NMAR—we tested the performance of case deletion, non-stochastic regression imputation, and multiple stochastic regression imputation with missing data rates of 10 percent, 20 percent, and 30 percent. Then we combined those results with the results for missing data rates of 5 percent and 40 percent to map out the relationship between the missing data rate and the performance of these three measures. We found that as the missing data rate increases, the bias also increased; however, these changes are smooth and gradual, revealing no obvious "tipping point."

References

Agodini, R., & M. Dynarski (2001). "Are Experiments the Only Option? A look at Dropout Prevention programs." ***The Review of Economics and Statistics,*** 86(1), 190-194.

Ahn, H. & J. L. Powell (1993). "Semiparametric Estimation of Censored Selection Models with a Nonparametric Selection Mechanism." ***Journal of Econometrics***, 58, 3-19.

Allison, P.D. (2002). ***Missing Data***. Thousand Oaks, CA: Sage University Paper No. 136.

Altonji, J. G., T. E. Elder, & C. R. Taber (2005). "Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools." ***Journal of Political Economy***. Vol. 113, February 2005.

Baker, S.G., G.M. Fitzmaurice, L.S. Freedman, & B.S. Kramer (2006) "Simple Adjustments For Randomized Trials With Nonramdomly Missing Or Censored Outcomes Arising From Informative Covariates." ***Biostatistics***, 7(1), 29-40.

Barnard, J., & X.L. Meng (1999), "Applications of Multiple Imputation in Medical Studies: From AIDS to NHANES." ***Statistical Methods in Medical Research***, 8, 17-36.

Battaglia, M. P., M.R. Frankel, & M. W. Link (2008). "Improving Standard Poststratification Techniques for Random-Digit-Dialing Telephone Surveys." ***Survey Research Methods***, 2(1), 11-19.

Bell, S.H., & L.L. Orr (1994). "Is subsidized employment cost effective for welfare recipients? Experimental evidence from seven state demonstrations." ***The Journal of Human Resources***, 29 (1), 42-61.

Bernstein, L., Dun Rappaport, C., Olsho, L., Hunt, D., & Levin, M. (2009). ***Impact Evaluation of the U.S. Department of Education's Student Mentoring Program*** (NCEE 2009-4047). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Bloom, H. (2005). "Randomizing Groups to Evaluate Place-Based Programs," in ***Learning More from Social Experiments***, New York: Russell Sage Foundation.

Bloom, H. (1984). "Accounting for no-shows in experimental evaluation designs." ***Evaluation Review***, 8(2), 225-246.

Bloom, H., et al., (2002). ***Can Non-Experimental Comparison Group Methods Match The Findings From A Random Assignment Evaluation Of Mandatory Work-to-Welfare Programs?*** New York, NY: MDRC.

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

Bryk, A. S. & S. W. Raudenbush (1992), ***Hierarchical Linear Models: Applications and Data Analysis Methods***. Newbury Park, CA: SAGE Publications.

Bullock, J. C. (2005). “Effects of the Accelerated Reader on reading performance of third, fourth, and fifth-grade students in one western Oregon elementary school.” University of Oregon; 0171 Advisor: Gerald Tindal. DAI, 66 (07A), 56-2529.

Burton, A. & D. G. Altman (2004). “Missing covariate data within cancer prognostic studies: a review of current reporting and proposed guidelines.” ***British Journal of Cancer***, 91, 4-8.

Campuzano, L., Dynarski, M., Agodini, R., & Rall, K. (2009). ***Effectiveness of Reading and Mathematics Software Products: Findings From Two Student Cohorts*** (NCEE 2009-4041). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Carpenter, J. R., & M. G. Kenward (2007). ***Missing Data in Randomised Control Trials—A Practical Guide***, unpublished monograph, http://www.missingdata.org.uk.

Chamberlain, G. (1986). “Asymptotic Efficiency in Semiparametric Models with Censoring.” ***Journal of Econometrics***, 32, 189-218.

Constantine, J., Player D., Silva, T., Hallgren, K., Grider, M., & Deke, J. (2009). ***An Evaluation of Teachers Trained Through Different Routes to Certification, Final Report*** (NCEE 2009-4043). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Cook, T.D., F. Habib, M. Phillips, R.A. Settersten, S.C. Shagle, & S.M. Degirmencioglu (1999). “Comer's School Development Program in Prince George's County, Maryland: A Theory-Based Evaluation.” ***American Educational Research Journal***, Vol. 36, No. 3, 543-597.

Corrin, W., Somers, M.-A., Kemple, J., Nelson, E., & Sepanik, S. (2008). ***The Enhanced Reading Opportunities Study: Findings from the Second Year of Implementation*** (NCEE 2009-4036). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Cox, B. (1991). “Weighting Survey Data for Analysis.” Paper presented at the American Statistical Association’s Continuing Education Program, Joint Statistical Meetings.

Das, M, W. Newey, & F Vella (2003). “Nonparametric Estimation of Sample Selection Models.” ***Review of Economic Studies***, 70, 33-58.

Dehejia, R., & S. Wahba (1999). “Causal Effects In Nonexperimental Studies: Reevaluating The Evaluation Of Training Programs.” ***Journal of the American Statistical Association***, 94 (448), 1053-1062.

Dehejia, R., & S. Wahba (2002). “Propensity Score Matching Methods For Nonexperimental Causal Studies.” ***Review of Economics and Statistics***, 84(1), 151-161.

DiNardo, J, J. McCrary, & L. Sanbonmatsu (2006). “Constructive Proposals for Dealing with Attrition: An Empirical Example.” NBER working paper, July 21, 2006. http://www-personal.umich.edu/~jdinardo/DMS_v9.pdf.

Food and Drug Administration (2006). “Guidance for Industry Patient-reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims (Draft Guidance).” Rockville, MD: U.S. Department of Health and Human Services, Food and Drug Administration.

Gamse, B.C., Jacob, R.T., Horst, M., Boulay, B., & Unlu, F. (2008). ***Reading First Impact Study Final Report*** (NCEE 2009-4038). Washington, DC: National Center for Education Evaluation.

Garet, M. S., S. Cronen, M. Eaton, A. Kurki, M. Ludwig, W. Jones, K. Uekawa, A. Falk, H. Bloom, F. Doolittle, P. Zhu, & L. Sztejnberg. ***The Impact of Two Professional Development Interventions on Early Reading Instruction and Achievement*** (NCEE 2008-4030). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Gennetian, L.A., P.A. Morris, J.M. Bos, & H.S. Bloom (2005). "Constructing Instrumental Variables From Experimental Data To Explore How Treatments Produce Effects." In ***Learning More from Social Experiments: Evolving Analytic Approaches***, Russell Sage Foundation, New York, H. Bloom editor.

Graham, J.W. (2009). "Missing Data Analysis: Making it Work in the Real World." ***Annual Review of Psychology***, 60, 549-576.

Haitovsky, Y. (1968). "Missing Data in Regression Analysis." ***Journal of the Royal Statistical Society***, Series B, 30, 67-82.

Heckman, J.J. (1976). "The Common Structure Of Statistical Models Of Truncated, Sample Selection And Limited Dependent Variables, And A Simple Estimator Of Such Models." ***Annals of Economic and Social Measurement***, 5, 475-492.

Heckman, J.J., & V.J. Hotz, V.J (1989). ***Choosing Among Alternative Nonexperimental Methods For Estimating The Impact Of Social Programs: The Case Of Manpower Training***. University of Chicago, Economics Research Center.

Horovitz, J.L. & C. F. Manski (1998). "Censoring of Outcomes and Regressors Due to Survey Nonresponse: Identification and Estimation Using Weights and Imputations." ***Journal of Econometrics***, 84, 37-58.

Horovitz, J.L. & C. F. Manski (2000). "Nonparametric Analysis of Randomized Experiments with Missing Covariate and Outcome Data." ***Journal of the American Statistical Association***, 95, 77-84.

Horton, N.J. & K.P. Kleinman (2007). "Much Ado About Nothing: A Comparison Of Missing Data Methods And Software To Fit Incomplete Data Regression Models." ***The American Statistician***, 61(1), 79-90.

Imbens, G.W. & J.M. Wooldridge (2009). "Recent Development in the Econometrics of Program Evaluation." ***Journal of Economic Literature***, 47(1), 5-86.

Horvitz, D.G. & Thompson, D.J. (1952). "A generalization of sampling without replacement from a finite universe." ***Journal of the American Statistical Association***, 47, 663-685.

Jones, M. (1996). "Indicator and Stratification Methods for Missing Explanatory Variables in Multiple Linear Regression." ***Journal of the American Statistical Association***, 91, 222-230.

Jones, M. (1996). "Indicator and Stratification Methods for Missing Explanatory Variables in Multiple Linear Regression." ***Journal of the American Statistical Association***, 91, 222-230.

Judkins, D., Hao, H., Barrett, B., & Adhikari, P. (2005). "Modeling and Polishing of Nonresponse Propensity." ASA Section on Survey Research Methods.

Klar, N., & A. Donner (2001). "Current and future challenges in the design and analysis of cluster randomized trials." ***Statistics in Medicine***, 20, 3729-3740.

Lessler, J.T. & W.D. Kalsbeek (1992). ***Nonsampling Errors in Surveys***. New York: Wiley.

Little R.J.A. (1986). "Survey Nonresponse Adjustments for Estimates of Means." ***International Statistical Review***, 54: 139-157.

Little, R.J.A. (1993). "Pattern-Mixture Models For Multivariate Incomplete Data." ***Journal of the American Statistical Association***, 88, 125-134.

Little, R.J.A. (1994). "A Class Of Pattern-Mixture Models For Normal Incomplete Data." Biometrika, 81, 471-483.

Little, R. & T. Raghunathan (2004). This reference is to a course named "Statistical Analysis with Missing Data," presented by Roderick Little and Trivellore Raghunathan, May 4-5, 2004, Arlington,Virginia.

Little, R.J.A., & D.B. Rubin (2002). ***Statistical Analysis with Missing Data***, 2nd Edition. Hoboken, NJ: John W. Wiley and Sons.

Lohr, S. L. (1999). ***Sampling: Design and Analysis***. Pacific Grove, CA: Duxbury Press.

Manski, C. F. (1990). "Nonparametric Bounds on Treatment Effects." ***American Economic Review, Papers and Proceedings***. 80: 319-323.

Mid-Atlantic Regional Education Laboratory (2008). "Research Design: The Effect of Connected Mathematics 2 (CM2) on the Math Achievement of Middle School Students in Selected Schools in the Mid-Atlantic Region: A Randomized Controlled Trial."

Moons, K.G.M., R.A.R.T. Donders, T. Stijnen, & F.E. Harrell (2006). "Using the outcome for imputation of missing predictor values was preferred." ***Journal of Clinical Epidemiology***, 59, 1092–1101.

Murray, D.M. (1998). ***Design and Analysis of Group-Randomized Trials***. New York: Oxford University Press.

Peugh, J.L., & C.K. Enders (2004), "Missing Data in Educational Research: A Review of Reporting Practices and Suggestions for Improvement." ***Review of Educational Research***, 74 (4), 525-556.

Powell, J. L. (1994). "Estimation of Semiparametric Models." ***Handbook of Econometrics, Volume IV*** (edited by R. F. Engle & D. McFadden). New York: North-Holland.

Price, C., B. Goodson, & G. Stewart (2007). ***Technical Report, Volume I: Infant Environmental Exposures and Neuropsychological Outcomes at Ages 7 to 10 Years.*** Prepared for National Immunization Program, Centers for Disease Control and Prevention, Atlanta GA.

Puma, M., S. Bell, R. Cook, C. Heid, & M. Lopez (2005). ***Head Start Impact Study: First Year Findings.*** US Department of Health and Human Service, Administration for Children and Families: Washington, DC.

Reiter, J.P. & Raghunathan, T.E. and Kinney, S.K. (2006). "The Importance of Modeling the Sampling Design in Multiple Imputation for Missing Data," ***Survey Methodology***, 32(2), p. 143.

Rosenbaum P.R. & D.B. Rubin (1983). "The Central Role of the Propensity Score in Observational Studies for Causal Effects." ***Biometrika***, 70, 41-55

Rosenbaum P.R. & D.B. Rubin (1984). "Reducing Bias in Observational Studies Using Subclassification on the Propensity Score." ***Journal of the American Statistical Association***, 79(387), 516-524.

Ross, S. M., J. Nunnery, & E. Goldfeder (2004). ***A randomized experiment on the effects of Accelerated Reader/Reading Renaissance in an urban school district: Preliminary evaluation report.*** Memphis, TN: The University of Memphis, Center for Research in Educational Policy.

Rubin, D.B. (1976), "Inference and Missing Data." ***Biometrika***, 63, 581-92.

Rubin, D.B. (1987). ***Multiple Imputation for Nonresponse in Surveys***. New York, NY: Wiley.

Rubin, D.B. (1996), "Multiple Imputation After 18+ Years," ***Journal of the American Statistical Association***, 91, 473 - 489.

SAS9 Documentation: http://support.sas.com/documentation/onlinedoc/91pdf/index.html.

SAS Institute Inc. (2003). ***SAS/STAT 9, 9.1 User's Guide***, Cary NC: SAS Institute, http://support.sas.com/onlinedoc/913/docMainpage.jsp.)

Schafer, J.L. (1999), "Multiple Imputation: A Primer," ***Statistical Methods in Medical Research***, 8, 3 - 15.

Shafer, J.L. (1997). ***Analysis of Incomplete Multivariate Data***. London: Chapman & Hall.

Schafer, J.L. & J.W. Graham (2002). "Missing Data: Our View of the State of the Art." ***Psychological Methods***, 7(2), 147-177.

Schochet, P.Z. (2005). ***Statistical Power for Random Assignment Evaluation of Education Programs.*** Mathematica Policy Research, Report Submitted to the Institute of Education Sciences, U.S. Department of Education.

Seftor, N.S., A. Mamun, & A. Schirm (2009). ***The Impacts of Regular Upward Bound on Postsecondary Outcomes 7-9 Years After Scheduled High School Graduation.*** Submitted by Mathematica Policy Research to the U.S. Department of Education, Policy and Program Studies Service.

Smith, J. P. Todd (2005). "Does Matching Overcome LaLonde's Critique of Nonexperimental Estimators?" ***Journal of Econometrics***. 125(1-2), 305-353.

Stolzenberg, R.M. & D.A. Relles (1990). "Theory Testing In A World Of Constrained Research Design: The Significance Of Heckmann's Censored Sampling Bias Correction For Nonexperimental Research." ***Sociological Methods & Research***, 18, 395-415.

Stolzenberg, R.M. & D.A. Relles (1997). "Tools For Intuition About Sample Selection Bias And Its Correction." American Sociological Review, 62, 494-507.

U.S. Department of Education (2009), "Identifying and Implementing Educational Practices Supported By Rigorous Evidence: A User Friendly Guide" http://www.ed.gov/rschstat/research/pubs/rigorousevid/guide_pg5.html

U.S. Department of Education (2008), ***What Works Clearinghouse, Procedures and Standards Handbook Version 2***, Washington: DC, December

van Buuren, S., H.C. Boshuizen, & D.L. Knook (1999), “Multiple Imputation of Missing Blood Pressure Covariates in Survival Analysis,” ***Statistics in Medicine***, 18, 681 - 694.

Wilde, E.T. & R. Hollister (2002). “How Close Is Close Enough? Testing Nonexperimental Estimates of Impact against Experimental Estimates of Impact with Education Test Scores as Outcomes.” Department of Economics, Swarthmore College.

Wolf, P., B. Gutmann, M. Puma, B. Kisida, L. Rizzo, & N. Eissa. ***Evaluation of the DC Opportunity Scholarship Program: Impacts After Three Years*** (NCEE 2009-4050). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Wood, A.M., I.R. White, & S.G. Thompson (2004), “Are Missing Outcome Data Adequatelt Handled? A Review of Published Randomized Controlled Trials in Major Medical Journals.” ***Clinical Trials,*** 1, 368-76.

Appendix A: Missing Data Bias as a Form of Omitted Variable Bias

One way to better understand the missing data problem is to see how it is related to the very first type of bias to which most of us were introduced in our first regression class, omitted variable bias. Suppose the true model of impacts is shown in equation (1):

(1) $Y = \beta_0 + \beta_1 X + \beta_2 Trt + \varepsilon$, where $\varepsilon \sim N(0, \sigma_1 \bullet I)$

where X is the baseline covariate and Trt is the treatment group indicator. However, suppose that the researchers conducting the RCT estimate a simpler model that excludes the baseline variable, as shown in equation (2):

(2) $Y = \alpha_0 + \alpha_1 Trt + v$, where $v \sim N(0, \sigma_2 \bullet I)$

The decision to estimate equation (2) instead of (1) may be driven by lack of knowledge—the researchers may not realize that the baseline variable affects the outcome—or by necessity—if the variable is inherently unobservable. However, the decision may have been based on the belief that "simpler is better" in RCTs since without missing data problems, RCTs yield unbiased impact estimates even when no control variables are included.

However, suppose some of the data are missing. More specifically, suppose that the outcome variable (Y) is missing for some cases, and the researchers plan to drop cases with missing values. (Other approaches to missing data are considered in the body of the report, but the consequences of dropping cases with missing values may be the best tool for illustrating the consequences of missing data.) In this very common scenario, will the researchers obtain unbiased estimates of the treatment effect (β_2)? The answer is "it depends" or "only in special cases." If the observations with non-missing values are just a *simple* random sample of the larger sample (MCAR), the answer would be yes. The only consequence is a smaller sample and less statistical power. If the observations with non-missing values are at least random *conditional the independent variables in the model* (the MAR category), then the answer is still yes.

What does this mean for our simple example? It means the RCT can obtain unbiased impact estimates if (1) the data are missing completely at random (just a coin toss or roll of the dice) or (2) the data are missing at random within each group defined by the only covariate included in the model: the treatment indicator. Exactly how this can be achieved is a core portion of the remainder of this appendix. Scenario (2) warrants some additional consideration since a difference in response rates between the treatment and control groups might be taken as a sign that the impact estimates are biased. However, **as long as the process behind the missing data is completely random within group,** it does not matter if the percentage of cases with missing data differs between the two groups: the treatment effect will still be unbiased.

However, there are still two potential pitfalls that could lead to biased estimates of the average impact of the treatment (both fall under the NMAR case). First, even where the occurrence of "missingness" is *unrelated* to treatment status, it can be *related* to other variables that have been omitted from the model (like X has been omitted from equation (2)) and cause bias. This is a case where missingness causes the observed treatment and control group outcome samples to be "equally unrepresentative" of the population of interest (i.e., the population these samples would represent if the outcomes data were totally complete). For example, suppose the outcome variable in equation (2) is the student's score on the state assessment in reading, and the observed baseline variable excluded from the model (X in equation (1)) equals 1 for Limited English Proficiency (LEP) students and 0 for non-LEP students. If the missing data rate is larger for LEP students than for other students, and equally larger for the treatment group and the control group samples, then the analysis sample of students—that is, the students with non-missing data—would be skewed toward non-LEP students in both the treatment group and the control group.

So when is this a problem? It is a problem *when the impact of the treatment differs between LEP students and other students*. For example, suppose the impact of the program on reading achievement is larger for LEP students than for other students. If LEP students are underrepresented in the analysis sample due to missing data, this will pull the estimated impacts downward. In this example, and many like it, random assignment will provide an unbiased estimate of the treatment's average impact *for students with nonmissing data*. However, because missing data has skewed both the treatment and control samples toward non-LEP students, for whom the impacts are relatively small, equation (2) will yield an downwardly biased estimate of the treatment's average impact *for students in the broader study sample* (and for whatever population this sample was designed to represent).

The second potential pitfall arises if missing data are related to *both* treatment status *and* a variable that has been omitted from the model. In this context, the analysis sample in both groups (treatment and control) will be unrepresentative of the broader population of students. However, because missing data is related to both treatment status and the omitted variable, the analysis samples in the treatment group and the control group will *not* be "equally unrepresentative," i.e., the treatment and control samples will be "differentially skewed" toward non-LEP students. While the first pitfall yields *unbiased* impact estimates for the wrong population, this pitfall yields *biased* impact estimates for the wrong population. In both instances, the wrong population is being studied in relation to the information policy makers need about the full set of students potentially impacted by an intervention.

To gain a better understanding of the second pitfall, let us build on the example developed in this section. The treatment and control samples used in the analysis could be "differentially skewed" toward non-LEP students if the treatment itself has a positive effect on English proficiency, and LEP students with higher English proficiency are more likely to be required to take the state test used to create the outcome variable for the analysis. In this scenario, within the analysis sample, the treatment group would be less skewed toward non-LEP students than the control group.

Mathematically, this introduces omitted variable bias by creating a positive correlation between treatment status (Trt in equation (2)) and the omitted variable (X or LEP status in equation (1)) in the observed sample.[78] Among students with complete data, LEP students are more likely to be in the treatment group than in the control group. If LEP status has a negative effect on the outcome—reading achievement, as measured by the state test—the positive correlation between the treatment and LEP status among students in the analysis sample will produce a negative bias in the impact estimate. Put differently, in this scenario, the RCT will understate the true impact of the treatment. More generally, when there are a variety of omitted variables that are related to both the outcome and its missing data pattern the bias due to missing data could be positive or negative.

There are two major lessons that can be gleaned from this discussion:

- **The situations under which we can obtain unbiased impact estimates if we simply exclude students with missing data are very restrictive**: Excluded students must be a random sample of students conditional on the independent variables included in the model, including the treatment indicator. Furthermore, the scenarios under which the missing data process is more complicated are quite plausible in many settings.
- **When missing data bias is considered, covariates play a more important role in RCTs than is commonly believed**. In fact, because we can see that bias due to missing data can be thought of as omitted variable bias, one approach to the problem is clear: include in the regressions used to estimate impacts variables that may influence both the outcome and the probability of having missing data on outcomes. While this is not the only approach to addressing missing data, it may be the simplest and most straightforward approach. Therefore, while covariates only serve to improve precision in RCTs when data are complete, in the real world, where data are never complete, covariates can help to reduce bias due to missing data.

[78] A non-zero correlation between the treatment/control group status indicator and an X variable measured prior to random assignment *can never occur for the sample as a whole* in an expected value sense, since treatment status is generated subsequent to that measurement and bears no relationship to anything else (having emerged from a random number generator or flip of a coin).

Appendix B: Resources for Using Multiple Imputation

In the section titled "Multiple Stochastic Regression Imputation," we provided some guidance on how to use multiple imputation to address missing data. Before implementing MI, or any other method to address missing data, we would recommend additional reading, such as Allison (2002) and articles by the statisticians who have developed and refined MI methods (e.g., Rubin, 1996; Schafer, 1999). However, in the end, researchers need to know how to use available software to implement MI should they choose that option for dealing with missing data. Therefore, we provide some guidance and references to other resources that may be helpful.

As shown earlier in this report, specialized software or MI-specific procedures in general purpose statistical software is not required to use MI methods. However, programming one's own multiple imputation algorithm is considerably more challenging than the programming required to specify analysis models in most evaluations. Therefore, specialized MI software may be useful for people who expect to conduct MI regularly. Furthermore, MI-specific procedures in the software that education researchers commonly use can make MI an easier choice in education-related RCTs.

In this section we list some specialized software packages for conducting MI, and we also list some MI-specific procedures in general purpose statistical software that may make MI easier for users to implement. For a comprehensive treatment of the software packages available to implement MI, see Horton & Kleinman (2007).[79] We conclude with a more extensive example of how to conduct MI in SAS for purposes of illustration. We have selected SAS for this example—without recommending it over other alternatives—because it is a commonly used general-purpose statistical package, and because it can handle the imputation, estimation, and combination steps all in a single package.

Software for Multiple Imputation

Specialized, stand-along software has been developed for implementing MI. Some examples include:

- **IVEware.** Developed by T. E. Raghunathan, Peter W. Solenberger, and John Van Hoewyk at the University of Michigan. It is available for download at www.isr.umich.edu/src/smp/ive/.
- **Amelia II.** Developed by James Honaker, Gary King, and Matthew Blackwell at Harvard University. It is available for download at http://gking.harvard.edu/amelia/.
- **SOLAS.** SOLAS is a commercial package that can be purchased at http://www.statsol.ie/html/solas/solas_home.html.

[79] This paper is available online at http://maven.smith.edu/~nhorton/muchado.pdf. The appendix showing code and output is available online at http://www.math.smith.edu/muchado-appendix.pdf.

Some statistical packages commonly used in education research also have MI procedures, modules, or options, while others do not. Some of the software packages used by education researchers include:

- **Stata.** A multiple imputation procedure developed by Patrick Royston can be installed directly through Stata.
- **SPSS.** SPSS Inc offers an add-on package named PASW Missing Values that will implement MI. The SPSS base package does not include canned routines for conducting MI.
- **HLM.** HLM can be used to analyze multiple data sets and can aggregate the results in an MI framework, provided that the multiple data sets are created by the user beforehand. http://www.ssicentral.com/hlm/example6-1.html
- **SPlus.** There are several Splus libraries available that contain functions for multiple imputation. These include:
  - Missing Data Library, built-in in Splus 6.0 and higher
  - Hmisc Library, for more information see http://www.multiple-imputation.com/
  - MICE. For more information see http://www.multiple-imputation.com/
  - NORM, CAT, MIX, and PAN. Developed by Joe Schafer at Penn State University. It is available for download at http://www.stat.psu.edu/~jls/misoftwa.html#top
- **R.** Most of the SPlus libraries listed above are also available for R. For more information, see http://cran.r-project.org/web/views/SocialSciences.html
- **SAS.** Specific SAS procedures have been developed to facilitate MI. See the example below.

An Example of MI Using SAS

SAS includes procedures that allow the user to (1) generate *k* multiple imputed values for each missing value in the data—which yields *k* different data sets—(2) estimate impacts for each imputed data set using one's preferred regression procedure (e.g., PROC MIXED for mixed, hierarchical, or multi-level modeling), and (3) combine the estimates across imputations. The last step will produce estimates of the coefficients in the model, including the treatment effect, and estimates of their standard errors.

Suppose we are conducting an RCT of an educational intervention, and 60 schools are randomly assigned—30 to treatment (T=1) and 30 to control (T=0). Furthermore, suppose that we want to estimate the average impacts of the intervention on three student outcomes, Y1, Y2, and Y3, controlling for four student-level background variables, X1, X2, X3, and X4, and two school-level descriptive variables, S1 and S2. The sample includes 1,000 students, but data for some students and some variables are missing. Suppose we plan to estimate impacts using a two-level model, where level 1 is the student-level model and level 2 is the school level model. Proc MI does not have the capability to explicitly fit at two-level "imputer's model", but we can approximate the two-level structure by adding 59 dummy variables corresponding to the 60 schools (less 1) to the imputers model. Let us represent those dummy variables as D1, D2, …, D59. We cannot simultaneously enter the school level variables S1 and S2 and the 59 dummy variables, so the variables S1 and S2 will not be used in the imputer's model, but their effects will be captured in the dummy variables.

In this context, MI can be used to address missing data in three steps:

Step 1 – Create Imputed Data

```
proc mi data=data1 noprint out=data2 seed=37851 NIMPUTE=5;
        var     T Y1 Y2 Y3 X1 X2 X3 X4  D1 - D59;
run;
```

One can use any number for the value of "seed." If we omit the seed value, SAS will generate are random number for use as the seed value. By explicitly specifying a seed value, as shown above, we can replicate our results if we re-run the same program at a later time. The seed's value does not matter; it is only a starting point for a procedure with a common end result using any seed.

This procedure reads the input data set *data1* and creates an output data set *data2* with 5 observations for every observation in *data1*. *Data2* contains a variable *_Imputation_* that equals 1, 2, 3, 4, or 5. Non-missing values for each variable are repeated across imputations; missing values are replaced with imputations based on a model that uses all of the variables in the *var* statement above.

Step 2 – Estimate the Model (e.g., Y1 only)

```
proc mixed data=data2;
        class school;  /* school is a variable that uniquely identifies each school
*/
        Model Y1 = T X1 X2 X3 S1 S2;
        by _Imputation_;
        random intercept/type=un sub=school;
        ods output SolutionF=data3a CovB=data3b;
run;
```

For each of the five imputed data sets, this procedure specifies a linear, multi-level model to estimate the average treatment effect on the first outcome variable (*Y1*). The *random* option allows the intercept to vary randomly across schools.

Step 3 – Combine the Estimates

```
proc mianalyze parms=data3a covb=data3b edf=994;  /* 994 = 1000 students –
6 X variables */
        var     T X1 X2 X3 S1 S2;
run;
```

This procedure combines the five sets of estimates. The output will include an estimate of the average treatment effect (coefficient on T) and its standard error.

Appendix C: Specifications for Missing Data Simulations

Introduction to Notation

The following notation is used throughout this appendix:

$Y_{\text{Pre,ij}}$ is a student achievement test score, measured at baseline (pre-treatment) for the i^{th} student, nested in the j^{th} school;

i = 1…60 (students per school); j = 1…60 (schools);

$Y_{\text{Post,ij}}$ is a student achievement test score, measured at follow-up (post-treatment) for the i^{th} student, nested in the j^{th} school;

$Female_{\text{ij}}$ = 1 if student is female, = 0 if male;

$Female\_Cen_{\text{ij}}$ is the grand-mean centered covariate for *Female*, obtained as

$$Female\_Cen_{\text{ij}} = Female_{\text{ij}} - \frac{\sum_{j=1}^{60}\sum_{i=1}^{60} Female_{\text{ij}}}{(60*60)}$$

$HiRisk_{\text{ij}}$ = 1 if student is high risk (e.g., low income), = 0 otherwise;

$HiRisk\_Cen_{\text{ij}}$ is the grand-mean centered covariate for *HiRisk*, obtained as

$$HiRisk\_Cen_{,\text{ij}} = HiRisk_{\text{ij}} - \frac{\sum_{j=1}^{60}\sum_{i=1}^{60} HiRisk_{\text{ij}}}{(60*60)}$$

Trt_{j} = 1 if school j was randomly assigned to the treatment condition, =0 if school j was randomly assigned to the control condition.

As part of the simulations, values of pretest and post-test scores were set to missing. The following variables represent the observed pretest and post-tests scores, where some of the scores are observed (non-missing) and others have missing values:

$Ymiss_{\text{Pre,ij}}$ is a pretest achievement score of the i^{th} student, nested in the j^{th} school; Some values are missing, others are non-missing.

$Ymiss_{\text{Post,ij}}$ is a post-test achievement score of the i^{th} student, nested in the j^{th} school; Some values are missing, others are non-missing.

Some additional notation is introduced in subsequent sections.

Hypothetical Education RCT Used in the Simulations

The assumed study design for the simulations is a randomized controlled trial (RCT) with random assignment of schools to treatment and control conditions. The goal of the fictional study that forms the basis of the simulations is to estimate the average impact of the treatment on student achievement. Key features of our fictional RCT design include: (1) 60 schools, with 30 assigned to treatment and 30 assigned to control; (2) 60 students per school; (3) baseline data on gender, an unspecified risk factor (e.g., low income), and pretest or pre-intervention achievement data in a single subject area (either reading or mathematics); and (4) follow-up outcome data on achievement in the same subject area as the pretest.

Estimation of the average impact of the hypothetical intervention on student achievement is assumed to be done using a 2-level hierarchical linear model, where students (level-1) are nested in schools (level-2), and the model includes gender and high risk status as student-level covariates. However, two different models are assumed to be estimated for the simulations: (1) Model A does *not* include a student-level pretest score as a covariate, and (2) Model B *does* include a student-level pretest score as a covariate:

Model A. Pretest score not available

$$Y_{\text{Post,ij}} = \beta_0 + \alpha_{0j} + \beta_1\,(Trt_j) + \beta_2\,(Female\_cen_{ij}) + \beta_3\,(HiRisk\_cen_{ij}) + \varepsilon_{ij}$$

Model B. Pretest score is available

$$Y_{\text{Post,ij}} = \beta_0 + \alpha_{0j} + \beta_1\,(Trt_j) + \beta_2\,(Female\_cen_{ij}) + \beta_3\,(HiRisk\_cen_{ij}) + \beta_4\,(Y_{\text{Pre,ij}}) + \varepsilon_{ij}$$

In each model, α_{0j} is a random school-level intercept that is assumed to be normally distributed with mean zero and variance τ^2, i.e., $\alpha_{0j} \sim N(0,\tau^2)$. It is also assumed to be independent of ε_{ij}, the student-level error term, and ε_{ij} is assumed to be normally distributed with mean 0 and variance σ^2, i.e., $\varepsilon_{ij} \sim N(0,\sigma^2)$. The coefficient $\hat{\beta}_1$ provides an estimate of the Intent-to-Treat Effect, or, in the absence of noncompliance, the average impact of the treatment.

Later in this appendix, when we describe how the different missing data methods are implemented, we will refer back to these two generic analysis models to indicate how we estimated the treatment effect when data were missing.

Generation of the Simulated Data

This section describes the generation of data for a single simulated data set. The process described here was replicated 1,000 times, producing 1,000 simulated data sets. Letting missing data occur at random (within defined probabilities) many times, and then averaging the results of estimation models across the 1,000 data sets, ensures the robustness of the simulation findings and of any conclusions about the performance of various missing data methodologies drawn from them. Multiple replications also provide *distributions* of impact estimates and their standard errors, reflective of the sampling variability built into the data (and present in real data). Estimates from these multiple replications converge on population parameters; for example, if there were no missing data and we increased the number of generated data sets towards infinity, the mean of the parameter estimates from the many simulations would converge to the true population mean. For scenarios where there are missing data, we use all the replications to determine the closeness of the impact estimator's mean across the replications to the true population parameter. This serves as the measure of bias in the impact estimate.

Generating Demographic Characteristics

To generate the sex and academic risk indicators, we first generated 60 school IDs, and within each school generated 60 student IDs. Within each school, we set the value of *Female* to "0" for 30 of the students, and set the value of *Female* to "1" for the remaining 30 students. Within each school, 12 students (6 females and 6 males) had the value of *HiRisk* set to "1," the remaining 48 students had the value of *HiRisk* set to "0." To summarize:

- each data set included 60 schools;
- each school consisted of 60 students—30 students (50%) were *Female,* and 12 students (20%) were *HiRisk* and,
- *Female* was independent of *HiRisk.*

Generating Pretest Scores

To generate pretest scores, we began by generating random errors terms for schools and students. To generate random school effects, we used the *Normal* function in SAS to generate 60 random normal deviates from a normal distribution with mean equal 0 and variance equal to 0.10. As will be shown subsequently, these will represent the deviations of each of the 60 school's individual intercepts from the grand mean intercept. In model notation, these are the values of α_{0j}, generated $\alpha_{0j} \sim N(0, \tau^2)$, where τ^2 is set to equal 0.10. In each simulated data set, each of the 60 schools was assigned one of these values. All students within a particular school shared the same common value on the school-level random deviate.

In the next step, we again used SAS's *Normal* function to generate values from a normal distribution. This time we generated 3,600 values from a distribution with mean equal 0 and variance equal to 0.90, corresponding to the 60 students within each of the 60 schools. Each of the simulated students was assigned a value from this random normal distribution. These values correspond to the random deviation terms, ε_{ij}, that represent the difference of an individual student's pretest score from his/her school's average value, and conditional on the student's covariate value. To summarize, we generated:

- school-level random effects, i.e., 60 values of α_{0j}, from a normal distribution with mean 0 and variance 0.10, and
- student-level random error terms, i.e., 3,600 values of ε_{ij}, from a normal distribution with mean zero and variance 0.90.

Next, we generated the values of each student's pretest (i.e., baseline) achievement score. The value of each student's pretest score was generated as a function of:

- a grand-mean intercept;
- student's gender;
- student's status on the *HiRisk* variable;

- the school-level mean pretest score, specifically, the school's deviation from the grand-mean intercept, α_{0j}; and,
- student-level residual error, ε_{ij}.

Using the values of the variables as described above, each student's pretest achievement score, $Y_{\text{Pre,ij}}$, was generated from the following equation:

$$Y_{\text{Pre,ij}} = \beta_0 + \beta_1(Female\_cen_{ij}) + \beta_2(HiRisk\_cen_{ij}) + \alpha_{0j} + \varepsilon_{ij}$$

where:

$$\begin{aligned}
&\beta_0 = 0\\
&\beta_1 = 0.20\\
&\beta_2 = -0.80\\
&\alpha_{0j} \sim N(0,0.1)\\
&\varepsilon_{ij} \sim N(0,0.9)
\end{aligned}$$

(See Chapter 4 for citations to justify our choices of β_1 and β_2.) Note that the mean of $Y_{\text{Pre,ij}}$ is,

$$\begin{aligned}
&= Mean(\beta_0) + Mean(\beta_1(Female\_cen_{ij})) + Mean(\beta_2(HiRisk\_cen_{ij})) + Mean(\alpha_{0j}) + Mean(\varepsilon_{ij})\\
&= 0 + (\beta_1)(0) + (\beta_2)(0) + 0 + 0\\
&= 0
\end{aligned}$$

And note that the level-1 (student-level) variance of $Y_{\text{Pre,ij}}$ is,

$$\begin{aligned}
&= Var(\beta_0) + Var(\beta_1(Female\_cen_{ij})) + Var(\beta_2(HiRisk\_cen_{ij})) + Var(\alpha_{0j}) + Var(\varepsilon_{ij})\\
&= 0 + (\beta_1)^2 Var(Female\_cen_{ij}) + (\beta_2)^2 Var(HiRisk\_cen_{ij}) + 0 + Var(\varepsilon_{ij})\\
&= 0 + (.2)^2(.5*.5) + (-.8)^2(.2*.8) + 0 + 0.90\\
&= 1.01
\end{aligned}$$

The level-2 (school-level) variance of $Y_{\text{Pre,ij}}$ is,

$$\begin{aligned}
&= Var(\alpha_{0j})\\
&= 0.10
\end{aligned}$$

Thus, the intraclass correlation (ICC) of the pretest scores is,

$$ICC = \frac{0.10}{(0.10 + 1.01)} = 0.09\,.$$

Generating Post-test Scores

To generate post-test scores, we began by generating random deviates for schools, $\alpha^*_{0j} \sim N(0, \tau^2)$, and students, $\varepsilon^*_{ij} \sim N(0, \sigma^2)$, where the stars are used to indicate that these are different sets than the random deviates used to create the pretest scores. The value of each student's post-test score was generated as a function of:

- a grand-mean intercept;
- student's gender;
- student's status on the *HiRisk* variable;
- student's pretest achievement score, $Y_{\mathrm{Pre,ij}}$;
- treatment status, Trt_{j};
- a negative interaction effect of treatment by pretest ($Trt_{\mathrm{j}} * Y_{\mathrm{Pre,ij}}$)—the treatment effect is larger for students with lower pretest scores than for students with higher pretest scores;
- the school-level mean post-test score, or put differently, the school's deviation from the grand-mean, α^*_{0j}; and,
- student-level residual error, ε^*_{ij}.

Post-test scores are generated from the following model:

$$Y_{\mathrm{Post,ij}} = \beta^*_0 + \beta^*_1(Female\_cen) + \beta^*_2(HiRisk\_cen) + \beta^*_3(Y_{\mathrm{Pre,ij}}) + \beta^*_4(Trt_j) + \beta^*_5(Trt_j * Y_{\mathrm{Pre,ij}}) + C(\alpha^*_{0j} + \varepsilon^*_{ij})$$

where:[80]

$$\beta^*_0 = 0$$
$$\beta^*_1 = 0.02$$
$$\beta^*_2 = -0.05$$
$$\beta^*_3 = \sqrt{0.50}$$
$$\beta^*_4 = 0.20$$
$$\beta^*_5 = -0.20/3$$
$$C = \sqrt{0.50}$$
$$\alpha^*_{0j} \sim N(0,0.1)$$
$$\varepsilon^*_{ij} \sim N(0,0.9)$$

[80] The constant C below was included as a multiplier in this equation to ensure that the unconditional variance of post-test scores would equal 1.

The process described above, to generate a single data set, was replicated 1,000 times to generate 1,000 data sets. Because we generated random values of α_{0j}, ε_{ij}, α^*_{0j}, and ε^*_{ij} from the distributions described above, each data set was different from all the others.

Missing Data Mechanism

The process described in the previous section was used to generate 1,000 complete data sets—that is, data sets without any missing values. In this section, we describe the process by which we generated missing values. Effectively, this involved selecting a random subsample from each of the 1,000 randomly generated samples, and for each subsample, setting the value of either the pretest or the post-test to missing.

The first step in this process involved specifying how the subsample would be selected. In particular, we specified the probability that the pretest or post-test would be missing from the data. Then these probabilities were used to select the subsample of cases for which the pretest or post-test would be set to missing.

In one set of simulations, we assumed that data were missing for a sample of students in each school. For these simulations, we randomly selected individual students within schools and set the value of their pretest score or post-test score to missing. In another set of simulations, we assumed that data were missing for entire schools (i.e., all students with a school had missing values). For these simulations, we randomly selected subsamples of schools and set the value of the pretest score or post-test score to missing for all students in these schools.

Missing Data Mechanism for Students

First, we generated missing values for those simulations in which data were missing for a sample of students in each school. We generated the missing indicator for three base scenarios--or missing data mechanisms—and for each base scenario, generated missing values such that either 5 percent of cases were missing or 40 percent of cases were missing.[81] For each combination, we also generated data in which the pretest was set to missing and data in which the post-test was set to missing. None of our simulations set both variables to missing simultaneously, and no other variables were set to missing (e.g., female or risk status) in any of our simulations.

Scenario I - Missingness depends on treatment assignment.

In this scenario, the missing data mechanism is dependent only on treatment assignment. In particular, the missing data rates are higher for students in control schools than for students in treatment schools. But within each group, missing cases are a simple random sample of all cases in the group.

Sub-scenario I, Missing data rate = low (5% overall)

For some simulations, we set either the pretest score or post-test score to missing for 5 percent of students in the sample. To do this, we first created a variable which indicated

[81] For a special analysis, we also generated missing data under three additional missing data rates—10 percent, 20 percent, and 30 percent , described subsequently in the section on Scenario III.

the probability of missing data, and we called it MissProb.. For treatment students (Trt = 1), MissProb was set to 0.04, and for control students (Trt = 0), MissProb was set to 0.06.

We then used SAS's *RanBin* function to generate values of 0 or 1 from a binomial distribution. The probability of generating a value of 1 was set to *MissProb*, and the new 0-1 variable was called *MissIndicator (e.g., MissIndicator* = *RanBin*(0,1, *MissProb)).*

Finally, we created values for the *observed* pretest scores and post-test scores, given that some values had been set to missing and would not be observed in the analysis. The observed pretest variable was set equal to the actual pretest score when the pretest score was non-missing; the observed pretest variable was set to a numeric missing value when the actual pretest score was missing from the data (e.g., not observed), as shown below:

- $Ymiss_{\text{Pre,ij}} = Y_{\text{Pre,ij}}$
- $Ymiss_{\text{Post,ij}} = Y_{\text{Post,ij}}$
- If *MissIndicator=1 then* $Ymiss_{\text{Pre,ij}}$ = . (set to missing)
- If *MissIndicator=1 then* $Ymiss_{\text{Post,ij}}$ = . (set to missing)

Sub-scenario II, Missing data rate = high (40% overall)

The missing data mechanism for this scenario was the same as that described in the previous section, except that the missing data rates were higher for both treatment schools and control schools. In particular:

- if *Trt* = 1 then *MissProb* = 0.35
- if *Trt* = 0 then *MissProb* = 0.45

All other steps were the same as described above.

Scenario II, Missingness depends on treatment assignment, pretest scores, and the interaction of the two.

The process we used for setting pretest and post-test scores to missing was the same as the process described for *Scenario I*, except that probability of missing values (*MissProb)* was dependent upon assignment to treatment, the pretest score, and the interaction of treatment assignment and pretest score. In particular:

- The missing data rate is higher in the control group than the treatment group;
- The missing data rate is higher for students with low pretest scores in both groups; but
- The relationship between pretest and the missing data rate is much stronger in the control group than in the treatment group (to generate a difference in the missing data mechanism between the two groups).

Sub-scenario I, missing data rate = low (5% overall)

For this subscenario, we set the missing data probability (MissProb) for each student as follows:

Quartile on Pretest Score	If *Trt*=1 Set *MissProb* to:	If *Trt=0* Set *MissProb* to:
4 (>= 0.717)	.03	.03
3 (>= 0.011, < 0.717)	.04	.05
2 (>= -0.703, < 0.011)	.04	.07
1 (< -0.703)	.05	.09
Overall Average	*.04*	*.06*

Note that in both treatment and control groups, the probability of missing data is higher for students with lower pretest scores, but that difference between the probabilities at the lowest and highest pretest quartiles is much greater in the control group than in the treatment group.

Sub-scenario II, missing data rate = high (40% overall)

For this subscenario, we set the missing data probabilities (MissProb) for each student as follows:

Quartile on Pretest Score	If *Trt*=1 Set *MissProb* to:	If *Trt=0* Set *MissProb* to:
4 (>= 0.717)	.30	.30
3 (>= 0.011, < 0.717)	.35	.40
2 (>= -0.703, < 0.011)	.35	.50
1 (< -0.703)	.40	.60
Overall Average	*.35*	*.45*

Scenario III, Missingness depends on treatment assignment, <u>post-test scores</u>, and the interaction of the two.

In particular:

- The missing data rate is higher in the control group than the treatment group;
- The missing data rate is higher for students with low post-test scores in both groups; but
- The relationship between post-test and the missing data rate is much stronger in the control group than in the treatment group (to generate a different missing data mechanism for the two groups).

Sub-scenario I, missing data rate = low (5% overall)

For this subscenario, we set the missing data probability (MissProb) for each student as follows:

If *Trt*=1		**If *Trt*=0**	
And Quartile on Post-Test Score is	**Then Set *MissProb* to:**	**And Quartile on Post-Test Score is**	**Then Set *MissProb* to:**
4 (>= 0.865)	.03	4 (>= 0.695)	.03
3 (>= 0.205, < 0.865)	.04	3 (>= 0.004, < 0.695)	.05
2 (>= -0.457, < 0.205)	.04	2 (>= -0.691, < 0.004)	.07
1 (< -0.457)	.05	1 (< -0.691)	.09
Overall Average	*.04*		*.06*

Note that for post-test scores, unlike pretest scores, the quartile cutoffs differ between the treatment and control groups due to the effect of the treatment on post-test scores.

Sub-scenario II, missing data rate = high (40% overall)

For this subscenario, we set the missing data probability (MissProb) for each student as follows:

If *Trt*=1		**If *Trt*=0**	
And Quartile on Post-Test Score is	**Then Set *MissProb* to:**	**And Quartile on Post-Test Score is**	**Then Set *MissProb* to:**
4 (>= 0.865)	.30	4 (>= 0.695)	.30
3 (>= 0.205, < 0.865)	.35	3 (>= 0.004, < 0.695)	.40
2 (>= -0.457, < 0.205)	.35	2 (>= -0.691, < 0.004)	.50
1 (< -0.457)	.40	1 (< -0.691)	.60
Overall Average	*.35*		*.45*

Other subscenarios

Under Scenario III, when the missing data mechanism depends on the post-test, we tested selected missing data methods at three different missing data rates between 5 percent and 40 percent: 10 percent, 20 percent, and 30 percent. The values of *MissProb* for these three missing data rates are provided below:

Missing Data Rate = 10 Percent			
If *Trt*=1		**If *Trt*=0**	
And Quartile on Post-Test Score is	**Then Set *MissProb* to:**	**And Quartile on Post-Test Score is**	**Then Set *MissProb* to:**
4 (>= 0.865)	.06	4 (>= 0.695)	.06
3 (>= 0.205, < 0.865)	.08	3 (>= 0.004, < 0.695)	.10
2 (>= -0.457, < 0.205)	.08	2 (>= -0.691, < 0.004)	.14
1 (< -0.457)	.10	1 (< -0.691)	.18
Overall Average	*.08*		*.12*

Missing Data Rate = 20 Percent			
If *Trt*=1		**If *Trt*=0**	
And Quartile on Post-Test Score is	**Then Set *MissProb* to:**	**And Quartile on Post-Test Score is**	**Then Set *MissProb* to:**
4 (>= 0.865)	.14	4 (>= 0.695)	.14
3 (>= 0.205, < 0.865)	.17	3 (>= 0.004, < 0.695)	.20
2 (>= -0.457, < 0.205)	.17	2 (>= -0.691, < 0.004)	.26
1 (< -0.457)	.20	1 (< -0.691)	.32
Overall Average	*.17*		*.23*

Missing Data Rate = 30 Percent			
If *Trt*=1		**If *Trt*=0**	
And Quartile on Post-Test Score is	**Then Set *MissProb* to:**	**And Quartile on Post-Test Score is**	**Then Set *MissProb* to:**
4 (>= 0.865)	.22	4 (>= 0.695)	.22
3 (>= 0.205, < 0.865)	.26	3 (>= 0.004, < 0.695)	.30
2 (>= -0.457, < 0.205)	.26	2 (>= -0.691, < 0.004)	.38
1 (< -0.457)	.30	1 (< -0.691)	.46
Overall Average	*.26*		*.34*

Missing Data Mechanism for Schools

In some RCTs, the missing data problem results from a lack of cooperation from schools and districts. Therefore, to account for this possibility, we ran a set of simulations under the assumption that data were missing for either 5 percent or 40 percent of *schools*—instead of for 5 percent or 40 percent of *students* within each school. The process used to generate missing values for all students in selected schools was largely parallel to the process used to generate missing values for selected students in each school.

However, for schools, the process for setting the missing data indicator to "1" operated at the school level. When school had a value of "1" on the missing data indicator, all pretest or all post-test scores within that school were set to missing. For example, for *Scenario 1*, the probability of missing data for an entire school was set to 4 percent for treatment schools and 6 percent for control schools. Within those schools—if picked as missing data cases---all pretest scores or all post-test scores were set to missing. For *Scenarios* II and III, quartiles were created from school-level means of pretest scores or post-test scores. However, the missing data probabilities were set to exactly the same values as shown in the previous section for missing students within schools.

Missing Data Methods

The following missing data methods were tested in the simulations under each of the missing data scenarios described in the previous section:

- Case deletion,
- Dummy variable adjustment,
- Mean value imputation,
- Non-stochastic regression imputation,
- Single stochastic regression imputation,
- Multiple stochastic regression imputation,
- Maximum Liklihood─EM algorithm with multiple imputation,
- Simple weighting,
- Sophisticated weighting, and,
- Fully-specified regression models with treatment/covariate interactions.

Case Deletion

Case deletion means simply that, if there is a missing value for any variable used in the model, the entire observation (student or school) is omitted from the analysis. This method is also known as complete case analysis because only observations that have complete data (no missing values) for every variable in the model are used in the analysis.

Therefore, regardless of whether we are missing pretest scores or post-test scores, and regardless of whether data are missing for students within schools or for entire schools, we implemented case deletion by dropping the cases with missing values.

To estimate the treatment effect once cases had been deleted, we estimated either Model A or Model B, as described in the Generic Analysis Plan presented earlier in this appendix. When pretest score was missing for a fraction of the sample, we estimated Model B. When post-test score was missing for a fraction of the sample, the model we estimated depended on whether pretest scores were available or unavailable:

- When pretest scores were available, we estimated Model B.
- When pretest scores were unavailable, we estimated Model A.

Dummy Variable Adjustment (Missing Pretest Scores Only)

The dummy variable adjustment required the creation of two new variables, $Y.dv_{\text{Pre,ij}}$, and $Dummy_{\text{ij}}$, defined as follows:

$$\begin{array}{llll} Y.dv_{\text{Pre,ij}} & = Ymiss_{\text{Pre,ij}} & \text{if } Ymiss_{\text{Pre,ij}} & \text{is non-missing} \\ & = 0 & \text{if } Ymiss_{\text{Pre,ij}} & \text{is missing} \end{array}$$

$$\begin{array}{llll} Dummy_{\text{ij}} & = 1 & \text{if } Ymiss_{\text{Pre,ij}} & \text{is missing} \\ & = 0 & \text{if } Ymiss_{\text{Pre,ij}} & \text{is non-missing} \end{array}$$

The analytical model used to estimate the treatment effect is similar to Model B, but the true value of the pretest is replaced by $Y.dv_{\text{Pre,ij}}$, and the dummy variable $Dummy_{\text{ij}}$ is added to the model, as shown below:

$$Y_{\text{Post,ij}} = \beta_0 + \alpha_{0j} + \beta_1 (Trt_j) + \beta_2 (Female\_cen_{ij}) + \beta_3 (HiRisk\_cen_{ij}) + \beta_4 (Y.dv_{\text{Pre,ij}}) + \beta_5 (Dummy_{\text{ij}}) + \varepsilon_{ij}$$

Mean Value Imputation

Missing Pretest Scores

When pretest scores are missing for a fraction of the sample, mean value imputation involves replacing the missing values of the pretest score with the mean of the non-missing values of the pretest score for students in the same group (treatment or control). The data were first divided into the two groups—the treatment group and the control group. In the treatment group, the variable $\overline{Ymiss}_{\text{Treat.Pre, ij}}$ was created as the mean of all non-missing values of$Ymiss_{\text{Pre,ij}}$. Similarly, for the control group, the variable $\overline{Ymiss}_{\text{Control.Pre,ij}}$ was created as the mean of all non-missing values of $Ymiss_{\text{Pre,ij}}$. Finally, the variable $Y.mv_{\text{Pre,ij}}$, was created as:

$$\begin{array}{llll} Y.mv_{\text{Pre,ij}} & = Ymiss_{\text{Pre,ij}} & \text{if } Ymiss_{\text{Pre,ij}} & \text{is non-missing} \\ & = \overline{Ymiss}_{\text{Treat.Pre, ij}} & \text{if } Ymiss_{\text{Pre,ij}} & \text{is missing and the student is in treatment group} \\ & = \overline{Ymiss}_{\text{Control.Pre,ij}} & \text{if } Ymiss_{\text{Pre,ij}} & \text{is missing and the student is in control group} \end{array}$$

The analytical model used to estimate the treatment effect is similar to Model B, but where the pretest variable with missing values is replaced by $Y.mv_{\text{Pre,ij}}$, as shown below:

$$Y_{\text{Post,ij}} = \beta_0 + \alpha_{0j} + \beta_1 (Trt_j) + \beta_2 (Female\_cen_{ij}) + \beta_3 (HiRisk\_cen_{ij}) + \beta_4 (Y.mv_{\text{Pre,ij}}) + \varepsilon_{ij}$$

Missing Post-test Scores

When post-test scores are missing for a fraction of the sample, mean value imputation is conducted just as we conducted it for missing pretest scores. For each group, treatment and control, we replaced the missing post-test values by the mean of the non-missing post-test score for students in the same group—that is, separately for the treatment and control groups—to create the outcome variable $Y.mv_{\text{Post,ij}}$.

For the simulations where we assumed pretest scores were available for the entire sample, the analytical model used to estimate the treatment effect is similar to Model B, but the true value of the post-test is replaced by $Y.mv_{\text{Post,ij}}$, as shown below:

$$Y.mv_{\text{Post,ij}} = \beta_0 + \alpha_{0j} + \beta_1 (Trt_j) + \beta_2 (Female\_cen_{ij}) + \beta_3 (HiRisk\_cen_{ij}) + \beta_4 (Y_{\text{Pre,ij}}) + \varepsilon_{ij}$$

For the simulations where we assumed pretest scores were not available for any sample members, the analytical model used to estimate the treatment effect is similar to Model A, but the post-test variable with missing values is replaced by $Y.mv_{\text{Post,ij}}$, as shown below:

$$Y.mv_{\text{Post,ij}} = \beta_0 + \alpha_{0j} + \beta_1 (Trt_j) + \beta_2 (Female\_cen_{ij}) + \beta_3 (HiRisk\_cen_{ij}) + \varepsilon_{ij}$$

Non-stochastic Regression Imputation

This method involves the replacement of missing values with predicted values from regression models. First we describe our approach to imputing values and analyzing the data when data were missing for students within schools; then we describe our approach to imputing values and analyzing the data when data are missing for entire schools.

Missing Pretest Scores for Students Within Schools

The data were first divided into the two groups—the treatment group and the control group. For the treatment group, we fit an imputer's model with the following form:

$$Ymiss_{\text{Pre,ij}} = \beta_0 + \beta_1 (Female\_cen_{ij}) + \beta_2 (HiRisk\_cen_{ij}) + \beta_3 (Y_{\text{Postij}}) + \sum_{j=1}^{29} \beta_{4+j} Sch_j + \varepsilon_{ij}$$

where Sch_j =1 if student is in school j, and = 0 else. Note the use of school fixed effects (e.g., school dummy variables) in this imputer's model instead of the random intercept terms for schools that are used in the analytical model used to estimate the treatment effect. This approach is consistent with the recommendations in Reiter, Raghunathan, and Kinney (2006).[82]

For treatment students, we obtained a predicted value of the pretest score as:

$$\hat{Y}_{\text{Treat.Pre,ij}} = \hat{\beta}_0 + \hat{\beta}_1 \,(Female\_cen_{ij}) + \hat{\beta}_2 \,(HiRisk\_cen_{ij}) + \hat{\beta}_3 \,(Y_{\text{Postij}}) + \sum_{j=1}^{29} \beta_{4+j} Sch_j$$

For control students, we estimated the same imputers model estimated for treatment students:

$$Ymiss_{\text{Pre,ij}} = \beta_0^* + \beta_1^* (Female\_cen_{ij}) + \beta_2^* (HiRisk\_cen_{ij}) + \beta_3^* (Y_{\text{Postij}}) + \sum_{j=30}^{59} \beta_{4+j}^* Sch_j + \varepsilon_{ij}^*$$

Then we used this model to produce predicted pretest scores for control students. Note that we put stars on parameters and estimates to emphasize that the model estimates for the control group are not identical to the model estimates for the treatment group:

$$\hat{Y}_{\text{Control.Pre,ij}} = \hat{\beta}_0^* + \hat{\beta}_1^* \,(Female\_cen_{ij}) + \hat{\beta}_2^* \,(HiRisk\_cen_{ij}) + \hat{\beta}_3^* \,(Y_{\text{Postij}}) \sum_{j=30}^{59} \hat{\beta}_{4+j}^* Sch_j$$

Finally, we created a new variable, $Y.nri_{\text{Pre,ij}}$, defined as follows:

$Y.nri_{\text{Pre,ij}}$	=	$Ymiss_{\text{Pre,ij}}$	if	$Ymiss_{\text{Pre,ij}}$	Is non-missing	
	=	$\hat{Y}_{\text{Treat.Pre,ij}}$	if	$Ymiss_{\text{Pre,ij}}$	is missing	and the student is in treatment group
	=	$\hat{Y}_{\text{Control.Pre,ij}}$	if	$Ymiss_{\text{Pre,ij}}$	is missing	and the student is in control group

The analytical model used to estimate the treatment effect is similar to Model B, but the pretest variable with missing values is replaced by $Y.nri_{\text{Pre,ij}}$, as shown below:

$$Y_{\text{Post,ij}} = \beta_0 + \alpha_{0j} + \beta_1 \,(Trt_j) + \beta_2 \,(Female\_cen_{ij}) + \beta_3 \,(HiRisk\_cen_{ij}) + \beta_4 \,(Y.nri_{\text{Pre,ij}}) + \varepsilon_{ij}$$

Missing Pretest Scores for Entire Schools

When schools had missing pretest scores (i.e., every student within a school had missing pretest values), we aggregated data to the school level, then used non-stochastic regression imputation to obtain predicted values of school-level mean pretest scores, then replaced the missing, school-level mean pretest scores with the imputed values, and then conducted impact analyses using the school-level aggregate data. We describe this process in more detail below.

[82] We conducted a set of simulations where the imputer's model included school random intercepts instead of school fixed effects. From this exercise, we found that the models with school fixed effects yielded more accurate standard error estimates than the models with school random effects. Therefore, in Section 4 and in Appendix E, we present results from the models that included school fixed effects.

For each school, the following school-level means were created from the observed (non-missing) student-level data:

$\overline{Ymiss_{\text{Pre.j}}}$ = the mean of $Ymiss_{\text{Pr}\,e.ij}$ over all students in school j (i.e., the school-level mean pretest score for schools with non-missing pretest scores)

$\overline{Y_{\text{Post.j}}}$ = the mean of $Y_{\text{Post.ij}}$ over all students in school j (i.e., school-level mean post-test score)

$\overline{Female\_cen_{.\text{j}}}$ = the mean of $Female\_Cen_{\text{ij}}$ over all students in school j (i.e., the proportion of students in the school who are female)

$\overline{HiRisk\_cen_{.\text{j}}}$ = the mean of $HiRisk\_Cen_{\text{ij}}$ over all students in school j (i.e., the proportion of students in the school who are at high-risk)

For the treatment group, we fit an imputer's model of the following form:

$$\overline{Ymiss_{\text{Pre.j}}} = \beta_0 + \beta_1 (\overline{Female\_cen_{.j}}) + \beta_2 (\overline{HiRisk\_cen_{.j}}) + \beta_3 (\overline{Y_{\text{Post.j}}}) + \varepsilon_j$$

Then we computed the predicted value from the regression for each school:

$$\hat{Y}_{\text{Treat.Pre.j}} = \hat{\beta}_0 + \hat{\beta}_1 (\overline{Female\_cen_{.j}}) + \hat{\beta}_2 (\overline{HiRisk\_cen_{.j}}) + \hat{\beta}_3 (\overline{Y_{\text{Post.j}}})$$

For the control group, we repeated the same steps. More specifically, we fit an imputer's model of the following form:

$$\overline{Ymiss_{\text{Pre.j}}} = \beta_0^* + \beta_1^* (\overline{Female\_cen_{.j}}) + \beta_2^* (\overline{HiRisk\_cen_{.j}}) + \beta_3^* (\overline{Y_{\text{Post.j}}}) + \varepsilon_j$$

The stars on the betas emphasize that the model estimates for the control group are not identical to the model estimates for the treatment group. For control schools, we computed the predicted value from the regression for each school:

$$\hat{Y}_{\text{Control.Pre.j}} = \hat{\beta}_0^* + \hat{\beta}_1^* (\overline{Female\_cen_{.j}}) + \hat{\beta}_2^* (\overline{HiRisk\_cen_{.j}}) + \hat{\beta}_3^* (\overline{Y_{Post.j}})$$

Finally, we created a new pretest variable, $\overline{Y.nri_{\text{Pre.j}}}$ as follows

$\overline{Y.nri_{\text{Pre.j}}}$ = $\overline{Ymiss_{\text{Pre.j}}}$ if $\overline{Ymiss_{\text{Pre.j}}}$ is non-missing

= $\hat{Y}_{\text{Treat.Pre.j}}$ if $\overline{Ymiss_{\text{Pre.j}}}$ is missing and the student is in treatment group

= $\hat{Y}_{\text{Control.Pre.j}}$ if $\overline{Ymiss_{\text{Pre.j}}}$ is missing and the student is in control group

The analytical model used to estimate the treatment effect is different from the Model B because the data has been aggregated to the school level. Therefore, we estimate a school-level analysis model, as shown below:

$$\overline{Y_{\text{Post.j}}} = \beta_0 + \beta_1 (Trt_j) + \beta_2 (\overline{Female\_cen_j}) + \beta_3 (\overline{HiRisk\_cen_j}) + \beta_4 (\overline{Y.nri_{\text{Pre.j}}}) + \varepsilon_j$$

Missing Post-test Scores for Students Within Schools

For missing post-test scores for students within schools, we took an almost identical approach to the imputation approach described earlier for addressing missing pretest scores for students within schools. However, instead of using the post-test to impute the pretest, we used the pretest to impute the post-test. The resulting outcome measure $Y.nri_{\text{Post,ij}}$ equals the true value when it is observed and the imputed value when the true value is missing.

For the simulations where we assumed pretest scores were available for the entire sample, the analytical model used to estimate the treatment effect is similar to Model B, but the post-test variable with missing values is replaced by $Y.nri_{\text{Post,ij}}$, as shown below:

$$Y.nri_{\text{Post,ij}} = \beta_0 + \alpha_{0j} + \beta_1 (Trt_j) + \beta_2 (Female\_cen_{ij}) + \beta_3 (HiRisk\_cen_{ij}) + \beta_4 (Y_{\text{Pre,ij}}) + \varepsilon_{ij}$$

For the simulations where we assumed pretest scores were not available for any sample members, the analytical model used to estimate the treatment effect is similar to Model A, but the post-test variable with missing values is replaced by $Y.nri_{\text{Post,ij}}$, as shown below:

$$Y.nri_{\text{Post,ij}} = \beta_0 + \alpha_{0j} + \beta_1 (Trt_j) + \beta_2 (Female\_cen_{ij}) + \beta_3 (HiRisk\_cen_{ij}) + \varepsilon_{ij}$$

Missing Post-test Scores for Entire Schools

For missing post-test scores for entire schools, we took an almost identical approach to the imputation approach described earlier for addressing missing pretest scores for entire schools—except that we used the school's mean pretest score to impute the school's mean post-test score, instead of the reverse.

When pretest data are available, the analytical model used to estimate the treatment effect is different from the Model B because the data has been aggregated to the school level. Therefore, we estimate a school-level analysis model, as shown below:

$$\overline{Y.nri_{\text{Post.j}}} = \beta_0 + \beta_1 (Trt_j) + \beta_2 (\overline{Female\_cen_j}) + \beta_3 (\overline{HiRisk\_cen_j}) + \beta_4 (\overline{Y_{\text{Pre.j}}}) + \varepsilon_j$$

When pretest data are not available, the analytical model used to estimate the treatment effect is different from the Model A—again because the data has been aggregated to the school level. Therefore, we estimate a school-level analysis model, as shown below:

$$\overline{Y.nri_{\text{Post.j}}} = \beta_0 + \beta_1 (Trt_j) + \beta_2 (\overline{Female\_cen_j}) + \beta_3 (\overline{HiRisk\_cen_j}) + \varepsilon_j$$

Single Stochastic Regression Imputation

Missing Pretest Scores for Students Within Schools

When pretest data are missing for students within schools, the procedure we used for implementing single stochastic regression imputation builds on the procedures we used for implementing non-stochastic regression imputation. However, in single stochastic regression imputation, a randomly selected residual is added to the predicted value from the imputer's model. For the treatment group, we fit the same imputer's model as for non-stochastic regression imputation; generate predicted values from the model, $\hat{Y}_{\text{Treat.Pre,ij}}$;

use the model to generate level-1 residuals, r_{ij}; and create a new outcome variable $Y.sri_{\text{Pre,ij}}$. This new outcome variable equals the true value when it is observed, and it equals $\hat{Y}_{\text{Treat.Pre,ij}} + r_{ij}$ when the true value is missing, where r_{ij} is a randomly selected residual. Finally, we repeat the process separately for the control group.

The analytical model used to estimate the treatment effect is similar to Model B, but the pretest variable with missing values is replaced by $Y.sri_{\text{Pre,ij}}$, as shown below:

$$Y_{\text{Post.ij}} = \beta_0 + \alpha_{0j} + \beta_1 (Trt_j) + \beta_2 (Female\_cen_{ij}) + \beta_3 (HiRisk\_cen_{ij}) + \beta_4 (Y.sri_{\text{Pre,ij}}) + \varepsilon_{ij}$$

Missing Pretest Scores for Entire Schools

When pretest data are missing for entire schools, the procedure for implementing single stochastic regression imputation is almost the same as that described earlier for non-stochastic regression imputation—except that a randomly selected residual is added to the predicted value. For the treatment group, we create a file of school-level means; fit the same school-level imputer's model; generate predicted values from the model for each treatment school, $\hat{Y}_{\text{Treat.Pre.j}}$; use the model to generate school-level residuals, r_j^k. We repeat this process for the control group to generate predicted values from the model for each control school, $\hat{Y}_{\text{Control.Pre.j}}$ and school-level residuals r_j^{l*}.

From these estimates, a new pretest variable is created as follows;

$\overline{Y.sri_{\text{Pre.j}}}$	=	$\overline{Ymiss_{\text{Pre.j}}}$	if	$\overline{Ymiss_{\text{Pre.j}}}$	is non-missing	
	=	$\hat{Y}_{\text{Treat.Pre.j}} + r_j^k$	if	$\overline{Ymiss_{\text{Pre.j}}}$	is missing	and student is in treatment group
	=	$\hat{Y}_{\text{Control.Pre.j}} + r_j^{l*}$	if	$\overline{Ymiss_{\text{Pre.j}}}$	is missing	and student is in control group

where r_j^k is a randomly selected residual from the treatment group, and r_j^{l*} is a randomly selected residual from the control group.

The analytical model used to estimate the treatment effect was of the form:

$$\overline{Y_{\text{Post.j}}} = \beta_0 + \beta_1 (Trt_j) + \beta_2 (\overline{Female\_cen_j}) + \beta_3 (\overline{HiRisk\_cen_j}) + \beta_4 (\overline{Y.sri_{\text{Pre.j}}}) + \varepsilon_j$$

Missing Post-test Scores for Students within Schools

When post-test data are missing for students within schools, the procedure we used for implementing single stochastic regression imputation builds on the procedures we used for implementing non-stochastic regression imputation. However, in single stochastic regression imputation, a randomly selected residual is added to the predicted value from the imputer's model. For the treatment group, we fit the same imputer's model as for non-stochastic regression imputation, which uses pretest scores to impute post-test scores; generate predicted values from the model, $\hat{Y}_{\text{Treat.Post,ij}}$; use the model to generate level-1 residuals, r_{ij}; and create a new outcome variable $Y.sri_{\text{Post,ij}}$. This new outcome variable equals the true value when it is observed, and it equals $\hat{Y}_{\text{Treat.Post,ij}} + r_{ij}$ when the true value is missing, for a randomly selected residual r_{ij}. Finally, we repeat the process separately for the control group.

For scenarios where pretest data were available, the analytical model used to estimate the treatment effect was of the form:

$$Y.sri_{\text{Post,ij}} = \beta_0 + \alpha_{0j} + \beta_1 (Trt_j) + \beta_2 (Female\_cen_{ij}) + \beta_3 (HiRisk\_cen_{ij}) + \beta_4 (Y_{\text{Pre,ij}}) + \varepsilon_{ij}$$

For scenarios where pretest data were not available, the analytical model used to estimate the treatment effect was of the form:

$$Y.sri_{\text{Post,ij}} = \beta_0 + \alpha_{0j} + \beta_1 (Trt_j) + \beta_2 (Female\_cen_{ij}) + \beta_3 (HiRisk\_cen_{ij}) + \varepsilon_{ij}$$

Missing Post-test Scores for Entire Schools

An analogous imputation procedure to that described for missing pretest scores of entire schools was used. To obtain imputed values for the treatment group, we create a file of school-level means; fit a school-level imputer's model to predict post-test school means; generate predicted values from the model for each treatment school; use the model to generate residuals; add residuals to predicted values to obtain imputed values; and replace missing values with imputed values. The process is repeated for the control group.

For scenarios where pretest data were available, the analytical model used to estimate treatment impact was of the form:

$$\overline{Y.sri_{\text{Post.j}}} = \beta_0 + \beta_1 (Trt_j) + \beta_2 (\overline{Female\_cen_j}) + \beta_3 (\overline{HiRisk\_cen_j}) + \beta_4 (\overline{Y_{\text{Pre.j}}}) + \varepsilon_j$$

For scenarios where pretest data were not available, the analytical model used to estimate treatment impact was of the form:

$$\overline{Y.sri_{\text{Post.j}}} = \beta_0 + \beta_1 (Trt_j) + \beta_2 (\overline{Female\_cen_j}) + \beta_3 (\overline{HiRisk\_cen_j}) + \varepsilon_j$$

Multiple Stochastic Regression Imputation

Multiple stochastic regression imputation is conducted in the same manner as single stochastic regression imputation, except that we produced five imputed values for each missing value.[83] Because each imputed value is created by randomly generating a residual and adding it to the imputation model's predicted value for the missing case, the five imputed values will be slightly different. Analysis of the five data sets produce five estimates of the treatment effect, which we will denote as $\hat{\beta}_1^1, \hat{\beta}_1^2, \hat{\beta}_1^3, \hat{\beta}_1^4, \hat{\beta}_1^5$. The overall treatment effect is computed as the mean of the five estimates. The standard error is computed as function of the standard error of each estimate, and the variation in the estimates across the five replications. For more details, see Chapter 3.

[83] The literature suggests that 5-10 imputations is adequate (see Rubin 1987, 1996 and Little & Rubin, 2002).

Missing Pretest Scores for Students Within Schools

For our implementation of multiple stochastic regression imputation, we used SAS's PROC MI to generate the imputed values, and SAS's PROC MIANALYZE to fit the analytical model to the data sets with the imputed values. One detail of our use of PROC MI to generate the imputed values is worthy of note. There is no way to fit a two-level hierarchical linear model (HLM) in PROC MI. Therefore, to approximate the two-level HLM model in our imputation model, we used fixed effects dummy variables for schools, in place of the random intercept terms that we used in the impact analysis models.

For the treatment group, we used PROC MI to fit an imputers model of the form:

$$Ymiss_{\text{Pre,ij}} = \beta_0 + \beta_1\,(Female\_cen_{ij}) + \beta_2\,(HiRisk\_cen_{ij}) + \beta_3\,(Y_{\text{Postij}}) + \sum_{j=1}^{29} Sch_j + \varepsilon_{ij}$$

where Sch_j =1 if student is in school j, and = 0 otherwise. We fit a model of the same form to the data from control group members. PROC MI then generates predicted values, and rather than sampling a residual, generates a residual from a normal distribution with mean 0 and variance equal to the estimated variance of ε_{ij}. The generated residual is added to the predicted value to obtain an imputed value, which we will denote as $\hat{Y}_{\text{Treat.Pre,ij}} + r_{ij}^{k}$ if student is in treatment group, and as $\hat{Y}_{\text{Control.Pre.j}} + r_j^{l*}$, if student is in control group.

As in single stochastic regression imputation, we define

$$\begin{array}{llllll} Y.mri_{\text{Pre,ij}} & = & Ymiss_{\text{Pre,ij}} & \text{if} & Ymiss_{\text{Pre,ij}} & \text{is non-missing} \\ & = & \hat{Y}_{\text{Treat.Pre,ij}} + r_{ij}^{k} & \text{if} & Ymiss_{\text{Pre,ij}} & \text{is missing and the student is in treatment group} \\ & = & \hat{Y}_{\text{Control.Pre,ij}} + r_{ij}^{l*} & \text{if} & Ymiss_{\text{Pre,ij}} & \text{is missing and the student is in control group} \end{array}$$

where r_j^k is a randomly selected residual from the set r_j, and r_j^{l*} is a randomly selected residual from the set r_j^*.

For each of the five data sets produced, the analytical model used to estimate the treatment effect was of the form:

$$Y_{\text{Post,ij}} = \beta_0 + \alpha_{0j} + \beta_1\,(Trt_j) + \beta_2\,(Female\_cen_{ij}) + \beta_3\,(HiRisk\_cen_{ij}) + \beta_4\,(Y.mri_{\text{Pre,ij}}) + \varepsilon_{ij}$$

The estimates from the five impact models were combined, as described in Chapter 3, to obtain the overall impact estimate and its standard error.

Missing Pretest Scores for Entire Schools

As described for non-stochastic and single stochastic regression imputation, data were aggregated to the school level to produce school-level means and imputation and impact analyses were conducted on the school-level data sets. We used SAS's PROC MI to fit the imputer's model, and SAS's PROC MIANALYZE to fit the analytical models to estimate impacts. The analytical model to estimate the treatment effect was of the form:

$$\overline{Y_{\text{Post.j}}} = \beta_0 + \beta_1\,(Trt_j) + \beta_2\,(\overline{Female\_cen_j}) + \beta_3\,(\overline{HiRisk\_cen_j}) + \beta_4\,(\overline{Y.mri_{\text{Pre.j}}}) + \varepsilon_j$$

The estimates from the five impact models were combined, as described in Chapter 3, to obtain the overall impact estimate and its standard error.

Missing Post-test Scores for Students Within Schools

An analogous imputation procedure to that described for missing pretest scores of students was used. For scenarios where pretest data were available, the analytical model used to estimate the treatment effect was of the form:

$$Y.mri_{\text{Post,ij}} = \beta_0 + \alpha_{0j} + \beta_1\,(Trt_j) + \beta_2\,(Female\_cen_{ij}) + \beta_3\,(HiRisk\_cen_{ij}) + \beta_4\,(Y_{\text{Pre,ij}}) + \varepsilon_{ij}$$

For scenarios where pretest data were not available, the analytical model used was of the form:

$$Y.mri_{\text{Post,ij}} = \beta_0 + \alpha_{0j} + \beta_1\,(Trt_j) + \beta_2\,(Female\_cen_{ij}) + \beta_3\,(HiRisk\_cen_{ij}) + \varepsilon_{ij}$$

The estimates from the five impact models were combined, as described in Chapter 3, to obtain the overall impact estimate and its standard error.

Missing Post-test Scores for Entire Schools

An analogous imputation procedure to that described for missing pretest scores fpr entire schools was used. As before, data were aggregated to the school level to produce school-level means, and imputation and the impact analyses were conducted on the school-level data sets. We used SAS's PROC MI to fit the imputer's model, and SAS's PROC MIANALYZE to fit the analytical models to estimate impacts. For scenarios where pretest data were available, the analytical model used was of the form:

$$\overline{Y.mri_{\text{Post.j}}} = \beta_0 + \beta_1\,(Trt_j) + \beta_2\,(\overline{Female\_cen_j}) + \beta_3\,(\overline{HiRisk\_cen_j}) + \beta_4\,(\overline{Y_{\text{Pre.j}}}) + \varepsilon_j$$

For scenarios where pretest data were not available, the analytical model used was of the form:

$$\overline{Y.mri_{\text{Post.j}}} = \beta_0 + \beta_1\,(Trt_j) + \beta_2\,(\overline{Female\_cen_j}) + \beta_3\,(\overline{HiRisk\_cen_j}) + \varepsilon_j$$

The estimates from the five impact models were combined, as described in Chapter 3, to obtain the overall impact estimate and its standard error.

Maximum Liklihood─EM Algorithm with Multiple Imputation

The EM algorithm with multiple imputation method was implemented in a manner very similar to that described for multiple stochastic regression imputation. The difference being that in the latter the imputed values were the predicted values from a regression model, and in the EM approach, the EM algorithm was used to obtain imputed values. In both approaches we generated five imputed data sets, and in both a random residual was added to each predicted value such that the imputed values in each of the five data sets would be slightly different from one another. Analysis of the five data sets produced five estimates of the treatment effect, which we denote as $\hat{\beta}_1^1, \hat{\beta}_1^2, \hat{\beta}_1^3, \hat{\beta}_1^4, \hat{\beta}_1^5$. The overall treatment effect is computed as the mean of the five estimates. The standard error is computed as function of the standard error of each estimate, and the variation in the estimates across the five replications. For more details, see Chapter 3.

Missing Pretest Scores for Students Within Schools

For our implementation of the EM algorithm with multiple imputation, we used SAS's PROC MI to generate the imputed values and SAS's PROC MIANALYZE to fit the analytical model to the data sets with the imputed values. As described previously, there is no way to specify the two-level hierarchical data structure (students nested in schools) in PROC MI. Therefore, to represent the two-level data structure in the implementation of the EM algorithm, we used fixed effects dummy variables for schools in place of the random intercept terms that we used in the impact analysis models.

For the treatment group, we entered the following variables into the EM algorithm:

$Ymiss_{\text{Pre,ij}}$

$Female\_cen_{ij}$

$HiRisk\_cen_{ij}$

Y_{Postij}

$Sch_1, Sch_2, \ldots, Sch_{29}$

where Sch_j =1 if student is in school j, and = 0 otherwise. We separately entered data for control group members into the EM algorithm. The same variables were entered, except the school dummies corresponded to the control group schools. PROC MI used the EM algorithm to generate predicted values and added a randomly generated residual to the predicted value to obtain an imputed value. We denote the imputed value as $\hat{Y}_{\text{Treat.Pre,ij}} + r_{ij}^k$ if student is in treatment group, and as $\hat{Y}_{\text{Control.Pre.j}} + r_j^{l*}$, if student is in control group.

As in multiple stochastic regression imputation, we define

$Y.emmi_{\text{Pre,ij}}$	=	$Ymiss_{\text{Pre,ij}}$	if	$Ymiss_{\text{Pre,ij}}$	is non-missing	
	=	$\hat{Y}_{\text{Treat.Pre,ij}} + r_{ij}^k$	if	$Ymiss_{\text{Pre,ij}}$	is missing	and the student is in treatment group
	=	$\hat{Y}_{\text{Control.Pre,ij}} + r_{ij}^{l*}$	if	$Ymiss_{\text{Pre,ij}}$	is missing	and the student is in control group

where r_j^k is a randomly selected residual from the set r_j, and r_j^{l*} is a randomly selected residual from the set r_j^*.

For each of the five data sets produced, the analytical model used to estimate the treatment effect was of the form:

$$Y_{\text{Post,ij}} = \beta_0 + \alpha_{0j} + \beta_1\,(Trt_j) + \beta_2\,(Female\_cen_{ij}) + \beta_3\,(HiRisk\_cen_{ij}) + \beta_4\,(Y.emmi_{\text{Pre,ij}}) + \varepsilon_{ij}$$

The estimates from the five impact models were combined, as described in Chapter 3, to obtain the overall impact estimate and its standard error.

Missing Pretest Scores for Entire Schools

As described for the regression imputation methods, data were aggregated to the school level to produce school-level means, and imputation and impact analyses were conducted on the school-level data sets. We used SAS's PROC MI to implement the EM algorithm, and SAS's PROC MIANALYZE to fit the analytical models to estimate impacts. The analytical model to estimate the treatment effect was of the form:

$$\overline{Y_{\text{Post.j}}} = \beta_0 + \alpha_{0j} + \beta_1\,(Trt_j) + \beta_2\,(\overline{Female\_cen_j}) + \beta_3\,(\overline{HiRisk\_cen_j}) + \beta_4\,(\overline{Y.emmi_{\text{Pre.j}}}) + \varepsilon_j$$

The estimates from the five impact models were combined, as described in Chapter 3, to obtain the overall impact estimate and its standard error.

Missing Post-test Scores for Students Within Schools

An analogous EM imputation procedure to that described for missing pretest scores of students was used. For scenarios where pretest data were available, the analytical model used to estimate the treatment effect was of the form:

$$Y.emmi_{\text{Post,ij}} = \beta_0 + \alpha_{0j} + \beta_1\,(Trt_j) + \beta_2\,(Female\_cen_{ij}) + \beta_3\,(HiRisk\_cen_{ij}) + \beta_4\,(Y_{\text{Pre,ij}}) + \varepsilon_{ij}$$

For scenarios where pretest data were not available, the analytical model used was of the form:

$$Y.emmi_{\text{Post,ij}} = \beta_0 + \alpha_{0j} + \beta_1\,(Trt_j) + \beta_2\,(Female\_cen_{ij}) + \beta_3\,(HiRisk\_cen_{ij}) + \varepsilon_{ij}$$

The estimates from the five impact models were combined, as described in Chapter 3, to obtain the overall impact estimate and its standard error.

Missing Post-test Scores for Entire Schools

An analogous EM imputation procedure to that described for missing pretest scores for entire schools was used. As before, data were aggregated to the school level to produce school-level means, and imputation and impact analyses were conducted on the school-level data sets. We used SAS's PROC MI to fit the imputer's model, and SAS's PROC MIANALYZE to fit the analytical models to estimate impacts. For scenarios where pretest data were available, the analytical model used was of the form:

$$\overline{Y.emmi_{\text{Post.j}}} = \beta_0 + \alpha_{0j} + \beta_1\,(Trt_j) + \beta_2\,(\overline{Female\_cen_j}) + \beta_3\,(\overline{HiRisk\_cen_j}) + \beta_4\,(\overline{Y_{\text{Pre.j}}}) + \varepsilon_j$$

For scenarios where pretest data were not available, the analytical model used was of the form:

$$\overline{Y.emmi}_{\text{Post.j}} = \beta_0 + \alpha_{0j} + \beta_1 (Trt_j) + \beta_2 (\overline{Female\_cen_j}) + \beta_3 (\overline{HiRisk\_cen_j}) + \varepsilon_j$$

The estimates from the five impact models were combined, as described in Chapter 3, to obtain the overall impact estimate and its standard error.

Fully-Specified Regression Models with Treatment/Covariate Interactions

Missing Post-test Scores for Students or Entire Schools

To implement this approach, we calculated the sample centered value of the pretest score by subtracting the sample mean of the pretest from the pretest score for each student:

$\overline{Y}^*_{\text{Pre,ij}}$= the mean of $Y_{\text{Pre,ij}}$

$YsampCen_{\text{Pre,ij}} = Y_{\text{Pre,ij}} - \overline{Y}^*_{\text{Pre,ij}}$. The mean of $YsampCen_{\text{Pre,ij}}$ is zero.

The analysis model is of the form:[84]

$$Ymiss_{\text{Post,ij}} = \beta_0 + \alpha_{0j} + \beta_1 (Trt_j) + \beta_2 (Female\_cen_{ij}) + \beta_3 (HiRisk\_cen_{ij}) +$$
$$\beta_4 (YsampCen_{\text{Pre,ij}}) + \beta_5 (Trt_j * YsampCen_{\text{Pre,ij}}) + \varepsilon_{ij}$$

and $\hat{\beta}_1$ is the estimate of the average treatment effect.[85]

Simple Weighting Approach

We use weighting to deal with missing post-test data only. Simple weighting can only be used when data are missing for students within schools, since it uses the non-missing cases in a school to represent the missing cases. With data missing for entire schools there are no non-missing to use.

Missing Post-test Scores for Students

For this method, in each school, respondents are simply weighted up to the total number of students sampled from the school.

Let N_j be the number of students sampled in the j^{th} school, and let n_j, be the number respondents in the the j^{th} school (i.e, the number of students with non-missing post-test scores). Within each school, each student with a non-missing post-test score is assigned a weight equal to $w_{ij} = \dfrac{N_j}{n_j}$; each student with a missing post-test score is assigned a weight of 0. Thus, for each school, the sum of the student weights will equal the number of students selected in the sample from that schools (N_j). For example, if 60 students were sampled in school *j* and 40 students had non-missing post-test scores, each of the 40

[84] This method assumes the pretest is available, so only one model is specified.

[85] Ordinarily, the fully specified regression model would have interaction terms between the treatment dummy variable and *all* of the baseline covariates in the model, not just some of them as shown here. Iinteractions of the treatment dummy and the female and risk covariates were not entered into the model here because in our synthetic data impact does not vary with these factors.

students would be assigned a weight equal to 60/40. The sum of the weights over the 40 respondents equals the size of the original sample.

Using the WEIGHT statement, the following models were fit to the data using SAS PROC MIXED.[86] For scenarios where pretest data were available, a weighted version of Model B was used to estimate the treatment effect, where the weight was set to w_{ij} (as defined in above). For scenarios where pretest data were not available, a weighted version of Model A was used to estimate the treatment effect, where the weight was set to w_{ij}.

More Sophisticated Weighting Approach

Missing Post-test Scores for Students or Entire Schools

Like the simple weighting approach described above, in this method we created weights for each respondent, then fit the same models as specified above to the complete cases, but applied the weights to the data using the weight statement in SAS PROC MIXED. The procedure for calculating the weights under the more sophisticated approach was as follows:

1. Estimate a logit model of response as a function of (1) dummy variables for 59 of the schools, (2) sex and race/ethnicity, and (3) pretest.
2. Using the model to compute estimated response probabilities for each student.
3. Divide the entire sample—including both respondents and nonrespondents—into quintiles based on the estimated response probability.
4. Compute the response rate (between 0 and 1) for each quintile.
5. Set the weight w_{ij} for each student to the inverse of the response rate for all students in the same quintile. This effectively creates five different weights—one for all students in each quintile: $w_1, w_2, w_3, w_4, and\ w_5$.

For scenarios where pretest data were available, a weighted version of Model B was used to estimate the treatment effect, where the weight was set to w_{ij}. For scenarios where pretest data were not available, a weighted version of Model A was used to estimate the treatment effect, where the weight was set to w_{ij}.

[86] In estimating the standard errors, we did not account either for the variation in the weights across sample members or for the sampling variability in the model estimates used to compute the weights. In principal, failure to account for these sources of variation should lead us to underestimate the standard error of the treatment effect. However, our simulation results suggest that the size of the bias in the estimated standard errors is very small (see third figure from each exhibit in Chapter 4). Therefore, while these corrections may be generally advisable with weighted data, we concluded that they were unnecessary for these simulations.

Appendix D: Full Set of Simulation Results

This appendix presents all tables of estimates produced in conducting the simulations reported in Chapter 4 and described in more detail in Appendix C. For these simulations, we generated 1,000 data sets, each with its own pattern of missing data . By letting missing data occur at random (within defined probabilities) many many times, and then averaging statistical results across the 1,000 data sets, we ensure the robustness of the simulation findings—and of the conclusions drawn from those findings concerning the performance of the different missing data methods examined. Multiple replications also give us *distributions* for the impact estimates and their standard errors, reflective of the sampling variability built into the data (and present in real data).

As described in Appendix C, different scenarios are used in the simulations, defined by (a) the nature of the missing data mechanism; (b) the missing data rate (5 percent or 40 percent); and (c) whether data are missing for students within schools or for entire schools. Therefore, the appendix contains 12 tables:

- Four tables for Scenario I (Table I.a.1 – Table I.b.2)
- Four tables for Scenario II (Table II.a.1 – Table II.b.2)
- Four tables for Scenario III (Table III.a.1 – Table III.b.2)

Each table consists of two panels:

1. Panel A, which shows the simulations results for situations where the pretest is missing for a fraction of the sample.
2. Panel B, which shows the simulations results for situations where the post-test is missing for a fraction of the sample.

The goal of these simulations was to estimate the bias in the impact estimates and standard errors from using different approaches to addressing missing data. Since bias is defined by the difference between the expected value of the estimator and the true parameter value, we estimated the bias in the two key estimates in the following way:

- **Impact estimate.** For the impact estimate, we estimated the bias by subtracting the true impact of 0.20 from the average of the impact estimates across the 1,000 samples.
- **Standard error.** For the estimate of the standard error of the impact estimate, we estimated the bias by subtracting an unbiased estimate of the standard error—the standard deviation of the 1,000 impact estimates—from the average of the standard error estimates across the 1,000 samples.

Note that each table begins by displaying the estimates from simulations in which none of the data were missing. These estimates do not match the true parameter values exactly due to random error. For example, the impact estimate with no missing data equals 0.203, which differs from the true impact of 0.200. When none of the data are missing, the impact estimates and standard error estimates are unbiased, and the non-zero bias estimates are entirely due to sampling error.

Table I.a.1: Scenario I, Missing Data Not Dependent on Pretest or Post-test: Data Missing for 5% of Students

		Impact Estimate				Standard Error of Impact Est.				90% CI
Data	**Pre-test Data Available?**	**Estimate**	**True Impact**	**Bias**	**Bias Level**	**Standard Estimate**	**Unbiased Estimate**	**Bias**	**Bias Level**	**% of Samples in Which 90% CI Contains .20**
No Missing Data	No	0.203	0.200	0.003	Low Bias	0.085	0.088	-0.002	Low Bias	0.892
	Yes	0.203	0.200	0.003	Low Bias	0.062	0.062	0.000	Low Bias	0.886
A. Pre-Test (X) Data Missing										
Case Deletion	No									
	Yes	0.201	0.200	0.001	Low Bias	0.062	0.063	-0.001	Low Bias	0.886
Dummy Variable Method	No									
	Yes	0.201	0.200	0.001	Low Bias	0.062	0.063	-0.001	Low Bias	0.886
Mean Value Imputation	No									
	Yes	0.201	0.200	0.001	Low Bias	0.062	0.063	-0.001	Low Bias	0.887
Single, Non-stochastic	No									
Regression Imputation	Yes	0.201	0.200	0.001	Low Bias	0.062	0.063	-0.001	Low Bias	0.882
Single, Stochastic	No									
Regression Imputation	Yes	0.201	0.200	0.001	Low Bias	0.062	0.063	-0.001	Low Bias	0.883
Multiple, Stochastic	No									
Regression Imputation (n = 5)	Yes	0.201	0.200	0.001	Low Bias	0.062	0.063	-0.001	Low Bias	0.883
EM Algorithm with Multiple Imputation	No									
(n = 5)	Yes	0.201	0.200	0.001	Low Bias	0.062	0.063	-0.001	Low Bias	0.884
B. Post-Test (Y) Data Missing										
Case Deletion	No	0.200	0.200	0.000	Low Bias	0.086	0.086	0.000	Low Bias	0.895
	Yes	0.201	0.200	0.001	Low Bias	0.062	0.063	-0.001	Low Bias	0.886
Mean Value Imputation	No	0.200	0.200	0.000	Low Bias	0.081	0.086	-0.004	Low Bias	0.876
	Yes	0.201	0.200	0.001	Low Bias	0.059	0.063	-0.004	Low Bias	0.868
Single, Non-stochastic	No	0.200	0.200	0.000	Low Bias	0.086	0.086	0.000	Low Bias	0.896
Regression Imputation	Yes	0.201	0.200	0.001	Low Bias	0.062	0.063	-0.001	Low Bias	0.886
Single, Stochastic	No	0.200	0.200	0.000	Low Bias	0.086	0.086	0.000	Low Bias	0.891
Regression Imputation	Yes	0.201	0.200	0.001	Low Bias	0.062	0.063	-0.001	Low Bias	0.883
Multiple, Stochastic	No	0.199	0.200	-0.001	Low Bias	0.086	0.086	0.000	Low Bias	0.898
Regression Imputation (n = 5)	Yes	0.200	0.200	0.000	Low Bias	0.062	0.063	-0.001	Low Bias	0.890
EM Algorithm with Multiple Imputation	No	0.200	0.200	0.000	Low Bias	0.086	0.086	0.001	Low Bias	0.899
(n = 5)	Yes									
Weighting - Simple	No	0.200	0.200	0.000	Low Bias	0.086	0.086	0.000	Low Bias	0.895
	Yes	0.201	0.200	0.001	Low Bias	0.062	0.063	-0.001	Low Bias	0.886
Weighting - Sophisticated	No	0.200	0.200	0.000	Low Bias	0.086	0.086	0.000	Low Bias	0.896
	Yes	0.201	0.200	0.001	Low Bias	0.062	0.063	-0.001	Low Bias	0.886
Fully Specified Regression Models	No									
w/ Treatment-Covariate Interactions	Yes	0.201	0.200	0.001	Low Bias	0.062	0.063	-0.001	Low Bias	0.888

Notes:
When pre-test scores are available, they are used as a covariate in the analysis model. In addition, we used pre-test scores to impute values and create weights. Bias estimates were computed as described in Chapter 4 and repeated at the beginning of this appendix. The level of the bias is characterized as "High Bias" or "Low Bias" based on the criteria established in Chapter 4. 90% CI refers to the 90-percent confidence interval around the impact estimate. For more details on the simulations, see Chapter 4 and Appendix C.

Table I.a.2: Scenario I, Missing Data Not Dependent on Pretest or Post-test: Data Missing for 5% of Schools

		Impact Estimate				Standard Error of Impact Est.				90% CI
Data	Pre-test Data Available?	Estimate	True Impact	Bias	Bias Level	Standard Estimate	Unbiased Estimate	Bias	Bias Level	% of Samples in Which 90% CI Contains .20
No Missing Data	No	0.203	0.200	0.003	Low Bias	0.085	0.088	-0.002	Low Bias	0.892
	Yes	0.203	0.200	0.003	Low Bias	0.062	0.062	0.000	Low Bias	0.886
A. Pre-Test (X) Data Missing										
Case Deletion	No									
	Yes	0.200	0.200	0.000	Low Bias	0.063	0.064	-0.001	Low Bias	0.895
Dummy Variable Method	No									
	Yes	0.200	0.200	0.000	Low Bias	0.063	0.064	-0.001	Low Bias	0.889
Mean Value Imputation	No									
	Yes	0.201	0.200	0.001	Low Bias	0.063	0.064	-0.001	Low Bias	0.887
Single, Non-stochastic	No									
Regression Imputation	Yes	0.200	0.200	0.000	Low Bias	0.061	0.064	-0.002	Low Bias	0.885
Single, Stochastic	No									
Regression Imputation	Yes	0.201	0.200	0.001	Low Bias	0.062	0.065	-0.002	Low Bias	0.886
Multiple, Stochastic	No									
Regression Imputation (n = 5)	Yes	0.200	0.200	0.000	Low Bias	0.063	0.064	-0.001	Low Bias	0.891
EM Algorithm with Multiple Imputation	No									
(n = 5)	Yes	0.200	0.200	0.000	Low Bias	0.063	0.064	-0.001	Low Bias	0.890
B. Post-Test (Y) Data Missing										
Case Deletion	No	0.199	0.200	-0.001	Low Bias	0.087	0.087	0.000	Low Bias	0.896
	Yes	0.200	0.200	0.000	Low Bias	0.063	0.064	-0.001	Low Bias	0.895
Mean Value Imputation	No	0.199	0.200	-0.001	Low Bias	0.083	0.087	-0.004	Low Bias	0.872
	Yes	0.200	0.200	0.000	Low Bias	0.061	0.066	-0.005	Low Bias	0.873
Single, Non-stochastic	No	0.199	0.200	-0.001	Low Bias	0.083	0.087	-0.004	Low Bias	0.872
Regression Imputation	Yes	0.200	0.200	0.000	Low Bias	0.061	0.064	-0.004	Low Bias	0.880
Single, Stochastic	No	0.198	0.200	-0.002	Low Bias	0.085	0.090	-0.005	Low Bias	0.871
Regression Imputation	Yes	0.199	0.200	-0.001	Low Bias	0.062	0.066	-0.004	Low Bias	0.876
Multiple, Stochastic	No	0.198	0.200	-0.002	Low Bias	0.087	0.087	0.000	Low Bias	0.892
Regression Imputation (n = 5)	Yes	0.199	0.200	-0.001	Low Bias	0.064	0.065	-0.001	Low Bias	0.894
EM Algorithm with Multiple Imputation	No	0.200	0.200	0.000	Low Bias	0.064	0.064	-0.001	Low Bias	0.891
(n = 5)	Yes									
Weighting - Simple	No									
	Yes									
Weighting - Sophisticated	No									
	Yes	0.200	0.200	0.000	Low Bias	0.063	0.064	-0.001	Low Bias	0.895
Fully Specified Regression Models	No									
w/ Treatment-Covariate Interactions	Yes	0.200	0.200	0.000	Low Bias	0.063	0.064	-0.001	Low Bias	0.895

Notes:
When pre-test scores are available, they are used as a covariate in the analysis model. In addition, we used pre-test scores to impute values and create weights. Bias estimates were computed as described in Chapter 4 and repeated at the beginning of this appendix. The level of the bias is characterized as "High Bias" or "Low Bias" based on the criteria established in Chapter 4. 90% CI refers to the 90-percent confidence interval around the impact estimate. For more details on the simulations, see Chapter 4 and Appendix C.

Table I.b.1 Scenario I, Missing Data Not Dependent on Pretest or Post-test: Data Missing for 40% of Students

		Impact Estimate				Standard Error of Impact Est.				90% CI
Data	Pre-test Data Available?	Estimate	True Impact	Bias	Bias Level	Standard Estimate	Unbiased Estimate	Bias	Bias Level	% of Samples in Which 90% CI Contains .20
No Missing Data	No	0.203	0.200	0.003	Low Bias	0.085	0.088	-0.002	Low Bias	0.892
	Yes	0.203	0.200	0.003	Low Bias	0.062	0.062	0.000	Low Bias	0.886
A. Pre-Test (X) Data Missing										
Case Deletion	No									
	Yes	0.203	0.200	0.003	Low Bias	0.065	0.065	0.000	Low Bias	0.888
Dummy Variable Method	No									
	Yes	0.203	0.200	0.003	Low Bias	0.069	0.069	-0.001	Low Bias	0.894
Mean Value Imputation	No									
	Yes	0.203	0.200	0.003	Low Bias	0.068	0.065	0.004	Low Bias	0.921
Single, Non-stochastic	No									
Regression Imputation	Yes	0.203	0.200	0.003	Low Bias	0.066	0.065	0.000	Low Bias	0.897
Single, Stochastic	No									
Regression Imputation	Yes	0.203	0.200	0.003	Low Bias	0.064	0.064	0.000	Low Bias	0.886
Multiple, Stochastic	No									
Regression Imputation (n = 5)	Yes	0.203	0.200	0.003	Low Bias	0.065	0.064	0.002	Low Bias	0.900
EM Algorithm with Multiple Imputation	No									
(n = 5)	Yes	0.202	0.200	0.002	Low Bias	0.066	0.064	0.002	Low Bias	0.900
B. Post-Test (Y) Data Missing										
Case Deletion	No	0.204	0.200	0.004	Low Bias	0.089	0.091	-0.002	Low Bias	0.887
	Yes	0.203	0.200	0.003	Low Bias	0.065	0.065	0.000	Low Bias	0.888
Mean Value Imputation	No	0.204	0.200	0.004	Low Bias	0.054	0.092	-0.038	High Bias	0.669
	Yes	0.203	0.200	0.003	Low Bias	0.041	0.073	-0.033	High Bias	0.650
Single, Non-stochastic	No	0.204	0.200	0.004	Low Bias	0.089	0.091	-0.002	Low Bias	0.891
Regression Imputation	Yes	0.203	0.200	0.003	Low Bias	0.065	0.065	0.000	Low Bias	0.889
Single, Stochastic	No	0.205	0.200	0.005	Low Bias	0.091	0.094	-0.003	Low Bias	0.885
Regression Imputation	Yes	0.204	0.200	0.004	Low Bias	0.066	0.067	-0.001	Low Bias	0.888
Multiple, Stochastic	No	0.204	0.200	0.004	Low Bias	0.095	0.092	0.003	Low Bias	0.911
Regression Imputation (n = 5)	Yes	0.203	0.200	0.003	Low Bias	0.069	0.065	0.004	Low Bias	0.905
EM Algorithm with Multiple Imputation	No	0.208	0.200	0.008	Low Bias	0.097	0.092	0.005	Low Bias	0.921
(n = 5)	Yes	0.204	0.200	0.004	Low Bias	0.070	0.066	0.005	Low Bias	0.904
Weighting - Simple	No	0.204	0.200	0.004	Low Bias	0.089	0.091	-0.002	Low Bias	0.889
	Yes	0.203	0.200	0.003	Low Bias	0.065	0.065	0.000	Low Bias	0.889
Weighting - Sophisticated	No	0.204	0.200	0.004	Low Bias	0.089	0.091	-0.002	Low Bias	0.890
	Yes	0.203	0.200	0.003	Low Bias	0.065	0.065	0.000	Low Bias	0.890
Fully Specified Regression Models	No									
w/ Treatment-Covariate Interactions	Yes	0.203	0.200	0.003	Low Bias	0.065	0.065	0.000	Low Bias	0.889

Notes:
When pre-test scores are available, they are used as a covariate in the analysis model. In addition, we used pre-test scores to impute values and create weights. Bias estimates were computed as described in Chapter 4 and repeated at the beginning of this appendix. The level of the bias is characterized as "High Bias" or "Low Bias" based on the criteria established in Chapter 4. 90% CI refers to the 90-percent confidence interval around the impact estimate. For more details on the simulations, see Chapter 4 and Appendix C.

Table I.b.2: Scenario I, Missing Data Not Dependent on Pretest or Post-test: Data Missing for 40% of Schools

		Impact Estimate				Standard Error of Impact Est.				90% CI
Data	Pre-test Data Available?	Estimate	True Impact	Bias	Bias Level	Standard Estimate	Unbiased Estimate	Bias	Bias Level	% of Samples in Which 90% CI Contains .20
No Missing Data	No	0.203	0.200	0.003	Low Bias	0.085	0.088	-0.002	Low Bias	0.892
	Yes	0.203	0.200	0.003	Low Bias	0.062	0.062	0.000	Low Bias	0.886
A. Pre-Test (X) Data Missing										
Case Deletion	No									
	Yes	0.202	0.200	0.002	Low Bias	0.081	0.080	0.001	Low Bias	0.901
Dummy Variable Method	No									
	Yes	0.203	0.200	0.003	Low Bias	0.073	0.075	-0.001	Low Bias	0.893
Mean Value Imputation	No									
	Yes	0.203	0.200	0.003	Low Bias	0.073	0.079	-0.006	Low Bias	0.869
Single, Non-stochastic	No									
Regression Imputation	Yes	0.201	0.200	0.001	Low Bias	0.055	0.081	-0.026	High Bias	0.736
Single, Stochastic	No									
Regression Imputation	Yes	0.201	0.200	0.001	Low Bias	0.062	0.079	-0.017	High Bias	0.780
Multiple, Stochastic	No									
Regression Imputation (n = 5)	Yes	0.205	0.200	0.005	Low Bias	0.069	0.074	-0.005	Low Bias	0.871
EM Algorithm with Multiple Imputation	No									
(n = 5)	Yes	0.200	0.200	0.000	Low Bias	0.072	0.076	-0.004	Low Bias	0.875
B. Post-Test (Y) Data Missing										
Case Deletion	No	0.202	0.200	0.002	Low Bias	0.112	0.112	0.000	Low Bias	0.886
	Yes	0.202	0.200	0.002	Low Bias	0.081	0.080	0.001	Low Bias	0.901
Mean Value Imputation	No	0.202	0.200	0.002	Low Bias	0.065	0.112	-0.047	High Bias	0.679
	Yes	0.202	0.200	0.002	Low Bias	0.055	0.097	-0.042	High Bias	0.677
Single, Non-stochastic	No	0.202	0.200	0.002	Low Bias	0.065	0.112	-0.047	High Bias	0.679
Regression Imputation	Yes	0.201	0.200	0.001	Low Bias	0.048	0.082	-0.034	High Bias	0.674
Single, Stochastic	No	0.205	0.200	0.005	Low Bias	0.083	0.129	-0.046	High Bias	0.694
Regression Imputation	Yes	0.203	0.200	0.003	Low Bias	0.060	0.095	-0.034	High Bias	0.702
Multiple, Stochastic	No	0.195	0.200	-0.005	Low Bias	0.107	0.114	-0.007	Low Bias	0.863
Regression Imputation (n = 5)	Yes	0.196	0.200	-0.004	Low Bias	0.077	0.084	-0.007	Low Bias	0.853
EM Algorithm with Multiple Imputation	No									
(n = 5)	Yes	0.204	0.200	0.004	Low Bias	0.085	0.083	0.002	Low Bias	0.878
Weighting - Simple	No									
	Yes									
Weighting - Sophisticated	No									
	Yes	0.202	0.200	0.002	Low Bias	0.081	0.080	0.001	Low Bias	0.899
Fully Specified Regression Models	No									
w/ Treatment-Covariate Interactions	Yes	0.202	0.200	0.002	Low Bias	0.081	0.080	0.001	Low Bias	0.900

Notes:
When pre-test scores are available, they are used as a covariate in the analysis model. In addition, we used pre-test scores to impute values and create weights. Bias estimates were computed as described in Chapter 4 and repeated at the beginning of this appendix. The level of the bias is characterized as "High Bias" or "Low Bias" based on the criteria established in Chapter 4. 90% CI refers to the 90-percent confidence interval around the impact estimate. For more details on the simulations, see Chapter 4 and Appendix C.

Table II.a.1: Scenario II, Missing Data Dependent on Pretest: Data Missing for 5% of Students

		Impact Estimate				Standard Error of Impact Est.				90% CI
Data	**Pre-test Data Available?**	**Estimate**	**True Impact**	**Bias**	**Bias Level**	**Standard Estimate**	**Unbiased Estimate**	**Bias**	**Bias Level**	**% of Samples in Which 90% CI Contains .20**
No Missing Data	No	0.203	0.200	0.003	Low Bias	0.085	0.088	-0.002	Low Bias	0.892
	Yes	0.203	0.200	0.003	Low Bias	0.062	0.062	0.000	Low Bias	0.886
A. Pre-Test (X) Data Missing										
Case Deletion	No									
	Yes	0.201	0.200	0.001	Low Bias	0.062	0.062	0.000	Low Bias	0.889
Dummy Variable Method	No									
	Yes	0.208	0.200	0.008	Low Bias	0.062	0.063	-0.001	Low Bias	0.884
Mean Value Imputation	No									
	Yes	0.213	0.200	0.013	Low Bias	0.062	0.063	-0.001	Low Bias	0.878
Single, Non-stochastic	No									
Regression Imputation	Yes	0.206	0.200	0.006	Low Bias	0.062	0.062	-0.001	Low Bias	0.883
Single, Stochastic	No									
Regression Imputation	Yes	0.206	0.200	0.006	Low Bias	0.062	0.062	0.000	Low Bias	0.881
Multiple, Stochastic	No									
Regression Imputation (n = 5)	Yes	0.207	0.200	0.007	Low Bias	0.062	0.062	-0.001	Low Bias	0.881
EM Algorithm with Multiple Imputation	No									
(n = 5)	Yes	0.206	0.200	0.006	Low Bias	0.062	0.062	0.000	Low Bias	0.884
B. Post-Test (Y) Data Missing										
Case Deletion	No	0.191	0.200	-0.009	Low Bias	0.085	0.089	-0.003	Low Bias	0.883
	Yes	0.201	0.200	0.001	Low Bias	0.062	0.062	0.000	Low Bias	0.889
Mean Value Imputation	No	0.189	0.200	-0.011	Low Bias	0.081	0.089	-0.008	Low Bias	0.864
	Yes	0.189	0.200	-0.011	Low Bias	0.059	0.063	-0.004	Low Bias	0.869
Single, Non-stochastic	No	0.191	0.200	-0.009	Low Bias	0.086	0.089	-0.003	Low Bias	0.882
Regression Imputation	Yes	0.202	0.200	0.002	Low Bias	0.062	0.062	0.000	Low Bias	0.890
Single, Stochastic	No	0.191	0.200	-0.009	Low Bias	0.086	0.089	-0.003	Low Bias	0.883
Regression Imputation	Yes	0.202	0.200	0.002	Low Bias	0.062	0.062	0.000	Low Bias	0.889
Multiple, Stochastic	No	0.191	0.200	-0.009	Low Bias	0.086	0.089	-0.003	Low Bias	0.884
Regression Imputation (n = 5)	Yes	0.201	0.200	0.001	Low Bias	0.062	0.062	0.000	Low Bias	0.893
EM Algorithm with Multiple Imputation	No	0.191	0.200	-0.009	Low Bias	0.086	0.089	-0.003	Low Bias	0.891
(n = 5)	Yes	0.201	0.200	0.001	Low Bias	0.062	0.062	0.000	Low Bias	0.890
Weighting - Simple	No	0.191	0.200	-0.009	Low Bias	0.086	0.089	-0.003	Low Bias	0.884
	Yes	0.201	0.200	0.001	Low Bias	0.062	0.062	0.000	Low Bias	0.891
Weighting - Sophisticated	No	0.191	0.200	-0.009	Low Bias	0.086	0.089	-0.003	Low Bias	0.884
	Yes	0.202	0.200	0.002	Low Bias	0.062	0.062	0.000	Low Bias	0.891
Fully Specified Regression Models	No									
w/ Treatment-Covariate Interactions	Yes	0.202	0.200	0.002	Low Bias	0.062	0.062	-0.001	Low Bias	0.891

Notes:
When pre-test scores are available, they are used as a covariate in the analysis model. In addition, we used pre-test scores to impute values and create weights. Bias estimates were computed as described in Chapter 4 and repeated at the beginning of this appendix. The level of the bias is characterized as "High Bias" or "Low Bias" based on the criteria established in Chapter 4. 90% CI refers to the 90-percent confidence interval around the impact estimate. For more details on the simulations, see Chapter 4 and Appendix C.

Table II.a.2: Scenario II, Missing Data Dependent on Pretest: Data Missing for 5% of Schools

		Impact Estimate				Standard Error of Impact Est.				90% CI
Data	**Pre-test Data Available?**	**Estimate**	**True Impact**	**Bias**	**Bias Level**	**Standard Estimate**	**Unbiased Estimate**	**Bias**	**Bias Level**	**% of Samples in Which 90% CI Contains .20**
No Missing Data	No	0.203	0.200	0.003	Low Bias	0.085	0.088	-0.002	Low Bias	0.892
	Yes	0.203	0.200	0.003	Low Bias	0.062	0.062	0.000	Low Bias	0.886
A. Pre-Test (X) Data Missing										
Case Deletion	No									
	Yes	0.201	0.200	0.001	Low Bias	0.063	0.064	-0.001	Low Bias	0.891
Dummy Variable Method	No									
	Yes	0.203	0.200	0.003	Low Bias	0.063	0.064	-0.001	Low Bias	0.885
Mean Value Imputation	No									
	Yes	0.205	0.200	0.005	Low Bias	0.063	0.064	-0.001	Low Bias	0.889
Single, Non-stochastic	No									
Regression Imputation	Yes	0.203	0.200	0.003	Low Bias	0.061	0.064	-0.003	Low Bias	0.874
Single, Stochastic	No									
Regression Imputation	Yes	0.204	0.200	0.004	Low Bias	0.062	0.064	-0.002	Low Bias	0.880
Multiple, Stochastic	No									
Regression Imputation (n = 5)	Yes	0.204	0.200	0.004	Low Bias	0.063	0.064	-0.002	Low Bias	0.883
EM Algorithm with Multiple Imputation	No									
(n = 5)	Yes	0.203	0.200	0.003	Low Bias	0.063	0.064	-0.002	Low Bias	0.884
B. Post-Test (Y) Data Missing										
Case Deletion	No	0.197	0.200	-0.003	Low Bias	0.087	0.091	-0.004	Low Bias	0.883
	Yes	0.201	0.200	0.001	Low Bias	0.063	0.064	-0.001	Low Bias	0.891
Mean Value Imputation	No	0.197	0.200	-0.003	Low Bias	0.083	0.091	-0.008	Low Bias	0.860
	Yes	0.198	0.200	-0.002	Low Bias	0.061	0.066	-0.005	Low Bias	0.877
Single, Non-stochastic	No	0.197	0.200	-0.003	Low Bias	0.083	0.091	-0.008	Low Bias	0.860
Regression Imputation	Yes	0.201	0.200	0.001	Low Bias	0.061	0.065	-0.005	Low Bias	0.866
Single, Stochastic	No	0.197	0.200	-0.003	Low Bias	0.085	0.092	-0.007	Low Bias	0.863
Regression Imputation	Yes	0.201	0.200	0.001	Low Bias	0.062	0.066	-0.004	Low Bias	0.869
Multiple, Stochastic	No	0.197	0.200	-0.003	Low Bias	0.087	0.091	-0.004	Low Bias	0.882
Regression Imputation (n = 5)	Yes	0.201	0.200	0.001	Low Bias	0.064	0.066	-0.002	Low Bias	0.884
EM Algorithm with Multiple Imputation	No									
(n = 5)	Yes	0.202	0.200	0.002	Low Bias	0.064	0.066	-0.002	Low Bias	0.882
Weighting - Simple	No									
	Yes									
Weighting - Sophisticated	No									
	Yes	0.202	0.200	0.002	Low Bias	0.063	0.064	-0.001	Low Bias	0.891
Fully Specified Regression Models	No									
w/ Treatment-Covariate Interactions	Yes	0.202	0.200	0.002	Low Bias	0.063	0.064	-0.001	Low Bias	0.892

Notes:
When pre-test scores are available, they are used as a covariate in the analysis model. In addition, we used pre-test scores to impute values and create weights. Bias estimates were computed as described in Chapter 4 and repeated at the beginning of this appendix. The level of the bias is characterized as "High Bias" or "Low Bias" based on the criteria established in Chapter 4. 90% CI refers to the 90-percent confidence interval around the impact estimate. For more details on the simulations, see Chapter 4 and Appendix C.

Table II.b.1: Scenario II, Missing Data Dependent on Pretest: Data Missing for 40% of Students

		Impact Estimate				Standard Error of Impact Est.				90% CI
Data	Pre-test Data Available?	Estimate	True Impact	Bias	Bias Level	Standard Estimate	Unbiased Estimate	Bias	Bias Level	% of Samples in Which 90% CI Contains .20
No Missing Data	No	0.203	0.200	0.003	Low Bias	0.085	0.088	-0.002	Low Bias	0.892
	Yes	0.203	0.200	0.003	Low Bias	0.062	0.062	0.000	Low Bias	0.886
A. Pre-Test (X) Data Missing										
Case Deletion	No									
	Yes	0.193	0.200	-0.007	Low Bias	0.065	0.065	0.000	Low Bias	0.887
Dummy Variable Method	No									
	Yes	0.238	0.200	0.038	Low Bias	0.068	0.069	-0.001	Low Bias	0.836
Mean Value Imputation	No									
	Yes	0.293	0.200	0.093	High Bias	0.068	0.065	0.004	Low Bias	0.622
Single, Non-stochastic	No									
Regression Imputation	Yes	0.252	0.200	0.052	High Bias	0.066	0.066	0.000	Low Bias	0.797
Single, Stochastic	No									
Regression Imputation	Yes	0.242	0.200	0.042	Low Bias	0.064	0.064	-0.001	Low Bias	0.821
Multiple, Stochastic	No									
Regression Imputation (n = 5)	Yes	0.242	0.200	0.042	Low Bias	0.065	0.063	0.002	Low Bias	0.834
EM Algorithm with Multiple Imputation	No									
(n = 5)	Yes	0.241	0.200	0.041	Low Bias	0.066	0.063	0.002	Low Bias	0.835
B. Post-Test (Y) Data Missing										
Case Deletion	No	0.110	0.200	-0.090	High Bias	0.088	0.090	-0.002	Low Bias	0.717
	Yes	0.193	0.200	-0.007	Low Bias	0.065	0.065	0.000	Low Bias	0.887
Mean Value Imputation	No	0.093	0.200	-0.107	High Bias	0.053	0.091	-0.038	High Bias	0.374
	Yes	0.093	0.200	-0.107	High Bias	0.041	0.072	-0.032	High Bias	0.270
Single, Non-stochastic	No	0.113	0.200	-0.087	High Bias	0.089	0.090	-0.002	Low Bias	0.724
Regression Imputation	Yes	0.201	0.200	0.001	Low Bias	0.065	0.065	0.000	Low Bias	0.887
Single, Stochastic	No	0.113	0.200	-0.087	High Bias	0.091	0.093	-0.002	Low Bias	0.727
Regression Imputation	Yes	0.201	0.200	0.001	Low Bias	0.066	0.066	0.000	Low Bias	0.894
Multiple, Stochastic	No	0.112	0.200	-0.088	High Bias	0.095	0.091	0.004	Low Bias	0.751
Regression Imputation (n = 5)	Yes	0.200	0.200	0.000	Low Bias	0.069	0.065	0.004	Low Bias	0.914
EM Algorithm with Multiple Imputation	No	0.112	0.200	-0.088	High Bias	0.096	0.091	0.005	Low Bias	0.761
(n = 5)	Yes	0.201	0.200	0.001	Low Bias	0.070	0.065	0.005	Low Bias	0.908
Weighting - Simple	No	0.112	0.200	-0.088	High Bias	0.089	0.090	-0.002	Low Bias	0.721
	Yes	0.193	0.200	-0.007	Low Bias	0.065	0.065	0.000	Low Bias	0.889
Weighting - Sophisticated	No	0.113	0.200	-0.087	High Bias	0.089	0.090	-0.002	Low Bias	0.724
	Yes	0.200	0.200	0.000	Low Bias	0.065	0.065	0.000	Low Bias	0.886
Fully Specified Regression Models	No									
w/ Treatment-Covariate Interactions	Yes	0.201	0.200	0.001	Low Bias	0.065	0.065	0.000	Low Bias	0.886

Notes:
When pre-test scores are available, they are used as a covariate in the analysis model. In addition, we used pre-test scores to impute values and create weights. Bias estimates were computed as described in Chapter 4 and repeated at the beginning of this appendix. The level of the bias is characterized as "High Bias" or "Low Bias" based on the criteria established in Chapter 4. 90% CI refers to the 90-percent confidence interval around the impact estimate. For more details on the simulations, see Chapter 4 and Appendix C.

Table II.b.2: Scenario II, Missing Data Dependent on Pretest: Data Missing for 40% of Schools

		Impact Estimate				Standard Error of Impact Est.				90% CI
Data	**Pre-test Data Available?**	**Estimate**	**True Impact**	**Bias**	**Bias Level**	**Standard Estimate**	**Unbiased Estimate**	**Bias**	**Bias Level**	**% of Samples in Which 90% CI Contains .20**
No Missing Data	No	0.203	0.200	0.003	Low Bias	0.085	0.088	-0.002	Low Bias	0.892
	Yes	0.203	0.200	0.003	Low Bias	0.062	0.062	0.000	Low Bias	0.886
A. Pre-Test (X) Data Missing										
Case Deletion	No									
	Yes	0.199	0.200	-0.001	Low Bias	0.081	0.082	-0.002	Low Bias	0.888
Dummy Variable Method	No									
	Yes	0.213	0.200	0.013	Low Bias	0.073	0.075	-0.002	Low Bias	0.890
Mean Value Imputation	No									
	Yes	0.233	0.200	0.033	Low Bias	0.073	0.078	-0.006	Low Bias	0.837
Single, Non-stochastic	No									
Regression Imputation	Yes	0.224	0.200	0.024	Low Bias	0.055	0.083	-0.029	High Bias	0.698
Single, Stochastic	No									
Regression Imputation	Yes	0.219	0.200	0.019	Low Bias	0.062	0.081	-0.019	High Bias	0.783
Multiple, Stochastic	No									
Regression Imputation (n = 5)	Yes	0.215	0.200	0.015	Low Bias	0.078	0.074	0.004	Low Bias	0.908
EM Algorithm with Multiple Imputation	No									
(n = 5)	Yes	0.222	0.200	0.022	Low Bias	0.074	0.076	-0.002	Low Bias	0.874
B. Post-Test (Y) Data Missing										
Case Deletion	No	0.165	0.200	-0.035	Low Bias	0.111	0.114	-0.003	Low Bias	0.873
	Yes	0.199	0.200	-0.001	Low Bias	0.081	0.082	-0.002	Low Bias	0.888
Mean Value Imputation	No	0.165	0.200	-0.035	Low Bias	0.064	0.114	-0.050	High Bias	0.622
	Yes	0.165	0.200	-0.035	Low Bias	0.055	0.099	-0.044	High Bias	0.615
Single, Non-stochastic	No	0.165	0.200	-0.035	Low Bias	0.064	0.114	-0.050	High Bias	0.622
Regression Imputation	Yes	0.202	0.200	0.002	Low Bias	0.048	0.086	-0.038	High Bias	0.651
Single, Stochastic	No	0.165	0.200	-0.035	Low Bias	0.083	0.120	-0.037	High Bias	0.710
Regression Imputation	Yes	0.200	0.200	0.000	Low Bias	0.061	0.090	-0.029	High Bias	0.729
Multiple, Stochastic	No	0.170	0.200	-0.030	Low Bias	0.152	0.121	0.031	Low Bias	0.929
Regression Imputation (n = 5)	Yes	0.205	0.200	0.005	Low Bias	0.112	0.092	0.020	Low Bias	0.932
EM Algorithm with Multiple Imputation	No									
(n = 5)	Yes	0.194	0.200	-0.006	Low Bias	0.087	0.089	-0.002	Low Bias	0.879
Weighting - Simple	No									
	Yes									
Weighting - Sophisticated	No									
	Yes	0.201	0.200	0.001	Low Bias	0.081	0.082	-0.002	Low Bias	0.891
Fully Specified Regression Models	No									
w/ Treatment-Covariate Interactions	Yes	0.202	0.200	0.002	Low Bias	0.081	0.082	-0.002	Low Bias	0.889

Notes:
When pre-test scores are available, they are used as a covariate in the analysis model. In addition, we used pre-test scores to impute values and create weights. Bias estimates were computed as described in Chapter 4 and repeated at the beginning of this appendix. The level of the bias is characterized as "High Bias" or "Low Bias" based on the criteria established in Chapter 4. 90% CI refers to the 90-percent confidence interval around the impact estimate. For more details on the simulations, see Chapter 4 and Appendix C.

Table III.a.1: Scenario III, Missing Data Dependent on Post-test: Data Missing for 5% of Students

		Impact Estimate				Standard Error of Impact Est.				90% CI
Data	**Pre-test Data Available?**	**Estimate**	**True Impact**	**Bias**	**Bias Level**	**Standard Estimate**	**Unbiased Estimate**	**Bias**	**Bias Level**	**% of Samples in Which 90% CI Contains .20**
No Missing Data	No	0.203	0.200	0.003	Low Bias	0.085	0.088	-0.002	Low Bias	0.892
	Yes	0.203	0.200	0.003	Low Bias	0.062	0.062	0.000	Low Bias	0.886
A. Pre-Test (X) Data Missing										
Case Deletion	No									
	Yes	0.195	0.200	-0.005	Low Bias	0.062	0.060	0.002	Low Bias	0.898
Dummy Variable Method	No									
	Yes	0.204	0.200	0.004	Low Bias	0.062	0.060	0.002	Low Bias	0.904
Mean Value Imputation	No									
	Yes	0.210	0.200	0.010	Low Bias	0.062	0.060	0.002	Low Bias	0.900
Single, Non-stochastic Regression Imputation	No									
	Yes	0.202	0.200	0.002	Low Bias	0.062	0.060	0.002	Low Bias	0.899
Single, Stochastic Regression Imputation	No									
	Yes	0.202	0.200	0.002	Low Bias	0.062	0.060	0.002	Low Bias	0.909
Multiple, Stochastic Regression Imputation (n = 5)	No									
	Yes	0.202	0.200	0.002	Low Bias	0.062	0.060	0.002	Low Bias	0.902
EM Algorithm with Multiple Imputation (n = 5)	No									
	Yes	0.202	0.200	0.002	Low Bias	0.062	0.060	0.002	Low Bias	0.904
B. Post-Test (Y) Data Missing										
Case Deletion	No	0.187	0.200	-0.013	Low Bias	0.086	0.084	0.001	Low Bias	0.895
	Yes	0.195	0.200	-0.005	Low Bias	0.062	0.060	0.002	Low Bias	0.898
Mean Value Imputation	No	0.185	0.200	-0.015	Low Bias	0.081	0.084	-0.003	Low Bias	0.876
	Yes	0.186	0.200	-0.014	Low Bias	0.059	0.060	-0.001	Low Bias	0.868
Single, Non-stochastic Regression Imputation	No	0.187	0.200	-0.013	Low Bias	0.086	0.084	0.001	Low Bias	0.894
	Yes	0.196	0.200	-0.004	Low Bias	0.062	0.060	0.002	Low Bias	0.898
Single, Stochastic Regression Imputation	No	0.187	0.200	-0.013	Low Bias	0.086	0.085	0.001	Low Bias	0.895
	Yes	0.195	0.200	-0.005	Low Bias	0.062	0.061	0.002	Low Bias	0.895
Multiple, Stochastic Regression Imputation (n = 5)	No	0.188	0.200	-0.012	Low Bias	0.086	0.084	0.002	Low Bias	0.898
	Yes	0.196	0.200	-0.004	Low Bias	0.062	0.060	0.002	Low Bias	0.904
EM Algorithm with Multiple Imputation (n = 5)	No	0.188	0.200	-0.012	Low Bias	0.086	0.084	0.002	Low Bias	0.902
	Yes	0.195	0.200	-0.005	Low Bias	0.062	0.060	0.002	Low Bias	0.900
Weighting - Simple	No	0.187	0.200	-0.013	Low Bias	0.086	0.084	0.001	Low Bias	0.894
	Yes	0.195	0.200	-0.005	Low Bias	0.062	0.060	0.002	Low Bias	0.898
Weighting - Sophisticated	No	0.187	0.200	-0.013	Low Bias	0.086	0.084	0.001	Low Bias	0.894
	Yes	0.195	0.200	-0.005	Low Bias	0.062	0.060	0.002	Low Bias	0.898
Fully Specified Regression Models w/ Treatment-Covariate Interactions	No									
	Yes	0.195	0.200	-0.005	Low Bias	0.062	0.060	0.002	Low Bias	0.898

Notes:
When pre-test scores are available, they are used as a covariate in the analysis model. In addition, we used pre-test scores to impute values and create weights. Bias estimates were computed as described in Chapter 4 and repeated at the beginning of this appendix. The level of the bias is characterized as "High Bias" or "Low Bias" based on the criteria established in Chapter 4. 90% CI refers to the 90-percent confidence interval around the impact estimate. For more details on the simulations, see Chapter 4 and Appendix C.

Table III.a.2: Scenario III, Missing Data Dependent on Post-test: Data Missing for 5% of Schools

		Impact Estimate				Standard Error of Impact Est.				90% CI
Data	Pre-test Data Available?	Estimate	True Impact	Bias	Bias Level	Standard Estimate	Unbiased Estimate	Bias	Bias Level	% of Samples in Which 90% CI Contains .20
No Missing Data	No	0.203	0.200	0.003	Low Bias	0.085	0.088	-0.002	Low Bias	0.892
	Yes	0.203	0.200	0.003	Low Bias	0.062	0.062	0.000	Low Bias	0.886
A. Pre-Test (X) Data Missing										
Case Deletion	No									
	Yes	0.199	0.200	-0.001	Low Bias	0.063	0.061	0.002	Low Bias	0.903
Dummy Variable Method	No									
	Yes	0.202	0.200	0.002	Low Bias	0.063	0.061	0.002	Low Bias	0.904
Mean Value Imputation	No									
	Yes	0.204	0.200	0.004	Low Bias	0.063	0.062	0.002	Low Bias	0.899
Single, Non-stochastic Regression Imputation	No									
	Yes	0.202	0.200	0.002	Low Bias	0.061	0.061	0.000	Low Bias	0.898
Single, Stochastic Regression Imputation	No									
	Yes	0.202	0.200	0.002	Low Bias	0.062	0.062	0.000	Low Bias	0.895
Multiple, Stochastic Regression Imputation (n = 5)	No									
	Yes	0.203	0.200	0.003	Low Bias	0.062	0.062	0.001	Low Bias	0.894
EM Algorithm with Multiple Imputation (n = 5)	No									
	Yes	0.202	0.200	0.002	Low Bias	0.063	0.061	0.002	Low Bias	0.905
B. Post-Test (Y) Data Missing										
Case Deletion	No	0.196	0.200	-0.004	Low Bias	0.088	0.086	0.002	Low Bias	0.902
	Yes	0.199	0.200	-0.001	Low Bias	0.063	0.061	0.002	Low Bias	0.903
Mean Value Imputation	No	0.196	0.200	-0.004	Low Bias	0.083	0.086	-0.003	Low Bias	0.881
	Yes	0.197	0.200	-0.003	Low Bias	0.061	0.062	-0.001	Low Bias	0.886
Single, Non-stochastic Regression Imputation	No	0.196	0.200	-0.004	Low Bias	0.083	0.086	-0.003	Low Bias	0.881
	Yes	0.200	0.200	0.000	Low Bias	0.061	0.061	-0.001	Low Bias	0.896
Single, Stochastic Regression Imputation	No	0.196	0.200	-0.004	Low Bias	0.085	0.088	-0.002	Low Bias	0.897
	Yes	0.199	0.200	-0.001	Low Bias	0.062	0.063	-0.001	Low Bias	0.894
Multiple, Stochastic Regression Imputation (n = 5)	No	0.193	0.200	-0.007	Low Bias	0.086	0.086	0.000	Low Bias	0.895
	Yes	0.198	0.200	-0.002	Low Bias	0.063	0.062	0.001	Low Bias	0.908
EM Algorithm with Multiple Imputation (n = 5)	No									
	Yes	0.200	0.200	0.000	Low Bias	0.064	0.062	0.002	Low Bias	0.909
Weighting - Simple	No									
	Yes									
Weighting - Sophisticated	No									
	Yes	0.199	0.200	-0.001	Low Bias	0.063	0.061	0.002	Low Bias	0.903
Fully Specified Regression Models w/ Treatment-Covariate Interactions	No									
	Yes	0.199	0.200	-0.001	Low Bias	0.063	0.061	0.002	Low Bias	0.902

Notes:
When pre-test scores are available, they are used as a covariate in the analysis model. In addition, we used pre-test scores to impute values and create weights. Bias estimates were computed as described in Chapter 4 and repeated at the beginning of this appendix. The level of the bias is characterized as "High Bias" or "Low Bias" based on the criteria established in Chapter 4. 90% CI refers to the 90-percent confidence interval around the impact estimate. For more details on the simulations, see Chapter 4 and Appendix C.

Table III.b.1: Scenario III, Missing Data Dependent on Post-test: Data Missing for 40% of Students

		Impact Estimate				Standard Error of Impact Est.				90% CI
Data	**Pre-test Data Available?**	**Estimate**	**True Impact**	**Bias**	**Bias Level**	**Standard Estimate**	**Unbiased Estimate**	**Bias**	**Bias Level**	**% of Samples in Which 90% CI Contains .20**
No Missing Data	No	0.203	0.200	0.003	Low Bias	0.085	0.088	-0.002	Low Bias	0.892
	Yes	0.203	0.200	0.003	Low Bias	0.062	0.062	0.000	Low Bias	0.886
A. Pre-Test (X) Data Missing										
Case Deletion	No									
	Yes	0.133	0.200	-0.067	High Bias	0.064	0.063	0.001	Low Bias	0.723
Dummy Variable Method	No									
	Yes	0.215	0.200	0.015	Low Bias	0.067	0.067	0.000	Low Bias	0.886
Mean Value Imputation	No									
	Yes	0.267	0.200	0.067	High Bias	0.069	0.065	0.003	Low Bias	0.749
Single, Non-stochastic	No									
Regression Imputation	Yes	0.200	0.200	0.000	Low Bias	0.065	0.064	0.001	Low Bias	0.899
Single, Stochastic	No									
Regression Imputation	Yes	0.201	0.200	0.001	Low Bias	0.064	0.063	0.001	Low Bias	0.890
Multiple, Stochastic	No									
Regression Imputation (n = 5)	Yes	0.201	0.200	0.001	Low Bias	0.065	0.062	0.003	Low Bias	0.903
EM Algorithm with Multiple Imputation	No									
(n = 5)	Yes	0.202	0.200	0.002	Low Bias	0.066	0.063	0.003	Low Bias	0.899
B. Post-Test (Y) Data Missing										
Case Deletion	No	0.076	0.200	-0.124	High Bias	0.087	0.087	0.000	Low Bias	0.574
	Yes	0.133	0.200	-0.067	High Bias	0.064	0.063	0.001	Low Bias	0.723
Mean Value Imputation	No	0.058	0.200	-0.142	High Bias	0.053	0.087	-0.035	High Bias	0.261
	Yes	0.057	0.200	-0.143	High Bias	0.040	0.069	-0.030	High Bias	0.127
Single, Non-stochastic	No	0.079	0.200	-0.121	High Bias	0.088	0.087	0.000	Low Bias	0.592
Regression Imputation	Yes	0.140	0.200	-0.060	High Bias	0.064	0.063	0.000	Low Bias	0.758
Single, Stochastic	No	0.080	0.200	-0.120	High Bias	0.090	0.089	0.001	Low Bias	0.600
Regression Imputation	Yes	0.141	0.200	-0.059	High Bias	0.065	0.065	0.001	Low Bias	0.763
Multiple, Stochastic	No	0.078	0.200	-0.122	High Bias	0.094	0.087	0.007	Low Bias	0.632
Regression Imputation (n = 5)	Yes	0.139	0.200	-0.061	High Bias	0.068	0.063	0.005	Low Bias	0.797
EM Algorithm with Multiple Imputation	No	0.080	0.200	-0.120	High Bias	0.095	0.088	0.007	Low Bias	0.636
(n = 5)	Yes	0.138	0.200	-0.062	High Bias	0.069	0.063	0.006	Low Bias	0.800
Weighting - Simple	No	0.079	0.200	-0.121	High Bias	0.088	0.087	0.000	Low Bias	0.588
	Yes	0.135	0.200	-0.065	High Bias	0.064	0.063	0.001	Low Bias	0.735
Weighting - Sophisticated	No	0.079	0.200	-0.121	High Bias	0.088	0.087	0.000	Low Bias	0.593
	Yes	0.140	0.200	-0.060	High Bias	0.064	0.063	0.001	Low Bias	0.761
Fully Specified Regression Models	No									
w/ Treatment-Covariate Interactions	Yes	0.138	0.200	-0.062	High Bias	0.064	0.063	0.001	Low Bias	0.752

Notes:
When pre-test scores are available, they are used as a covariate in the analysis model. In addition, we used pre-test scores to impute values and create weights. Bias estimates were computed as described in Chapter 4 and repeated at the beginning of this appendix. The level of the bias is characterized as "High Bias" or "Low Bias" based on the criteria established in Chapter 4. 90% CI refers to the 90-percent confidence interval around the impact estimate. For more details on the simulations, see Chapter 4 and Appendix C.

Table III.b.2: Scenario III, Missing Data Dependent on Post-test: Data Missing for 40% of Schools

		Impact Estimate				Standard Error of Impact Est.				90% CI
Data	**Pre-test Data Available?**	**Estimate**	**True Impact**	**Bias**	**Bias Level**	**Standard Estimate**	**Unbiased Estimate**	**Bias**	**Bias Level**	**% of Samples in Which 90% CI Contains .20**
No Missing Data	No	0.203	0.200	0.003	Low Bias	0.085	0.088	-0.002	Low Bias	0.892
	Yes	0.203	0.200	0.003	Low Bias	0.062	0.062	0.000	Low Bias	0.886
A. Pre-Test (X) Data Missing										
Case Deletion	No									
	Yes	0.176	0.200	-0.024	Low Bias	0.080	0.081	0.000	Low Bias	0.876
Dummy Variable Method	No									
	Yes	0.207	0.200	0.007	Low Bias	0.072	0.073	0.000	Low Bias	0.904
Mean Value Imputation	No									
	Yes	0.225	0.200	0.025	Low Bias	0.073	0.081	-0.008	Low Bias	0.851
Single, Non-stochastic	No									
Regression Imputation	Yes	0.203	0.200	0.003	Low Bias	0.055	0.081	-0.026	High Bias	0.734
Single, Stochastic	No									
Regression Imputation	Yes	0.203	0.200	0.003	Low Bias	0.062	0.078	-0.017	High Bias	0.790
Multiple, Stochastic	No									
Regression Imputation (n = 5)	Yes	0.197	0.200	-0.003	Low Bias	0.067	0.073	-0.006	Low Bias	0.865
EM Algorithm with Multiple Imputation	No									
(n = 5)	Yes	0.202	0.200	0.002	Low Bias	0.074	0.075	-0.001	Low Bias	0.897
B. Post-Test (Y) Data Missing										
Case Deletion	No	0.152	0.200	-0.048	Low Bias	0.110	0.112	-0.003	Low Bias	0.850
	Yes	0.176	0.200	-0.024	Low Bias	0.080	0.081	0.000	Low Bias	0.876
Mean Value Imputation	No	0.152	0.200	-0.048	Low Bias	0.064	0.112	-0.048	High Bias	0.590
	Yes	0.151	0.200	-0.049	Low Bias	0.055	0.099	-0.044	High Bias	0.573
Single, Non-stochastic	No	0.152	0.200	-0.048	Low Bias	0.064	0.112	-0.048	High Bias	0.590
Regression Imputation	Yes	0.177	0.200	-0.023	Low Bias	0.048	0.083	-0.036	High Bias	0.632
Single, Stochastic	No	0.150	0.200	-0.050	High Bias	0.082	0.126	-0.045	High Bias	0.655
Regression Imputation	Yes	0.176	0.200	-0.024	Low Bias	0.060	0.092	-0.032	High Bias	0.697
Multiple, Stochastic	No	0.166	0.200	-0.034	Low Bias	0.101	0.116	-0.015	Low Bias	0.807
Regression Imputation (n = 5)	Yes	0.187	0.200	-0.013	Low Bias	0.074	0.085	-0.011	Low Bias	0.824
EM Algorithm with Multiple Imputation	No									
(n = 5)	Yes	0.177	0.2	-0.023	Low Bias	0.086	0.085	0.001	Low Bias	0.876
Weighting - Simple	No									
	Yes									
Weighting - Sophisticated	No									
	Yes	0.177	0.200	-0.023	Low Bias	0.080	0.081	0.000	Low Bias	0.879
Fully Specified Regression Models	No									
w/ Treatment-Covariate Interactions	Yes	0.178	0.200	-0.022	Low Bias	0.080	0.081	0.000	Low Bias	0.877

Notes:
When pre-test scores are available, they are used as a covariate in the analysis model. In addition, we used pre-test scores to impute values and create weights. Bias estimates were computed as described in Chapter 4 and repeated at the beginning of this appendix. The level of the bias is characterized as "High Bias" or "Low Bias" based on the criteria established in Chapter 4. 90% CI refers to the 90-percent confidence interval around the impact estimate. For more details on the simulations, see Chapter 4 and Appendix C.

Appendix E: Standards for Judging the Magnitude of the Bias for Different Missing Data Methods

To assess the performance of different missing data methods, we set objective standards and applied those standards in interpreting the results. This appendix describes these standards and how they were chosen. In summary:

- We relied on the attrition standards of the *What Works Clearinghouse*.
- An impact estimate was considered to have "high bias" if the absolute value of the bias was greater than 0.05 of a standard deviation of the outcome measure.
- A standard error estimate was considered to have "high bias" if it yielded as much bias in the t-statistic as did bias in the impact estimate of 0.05 standard deviations.

A. Bias Standards for the Impact Estimates

To summarize the performance of the different measures, we identify the methods that yielded bias that would be considered "high" relative to a benchmark set by the *What Works Clearinghouse* (WWC). In developing its attrition standards, as well as its standards for baseline equivalence, the WWC decided that bias in the impact estimate of more than 0.05 standard deviations was unacceptably large (US ED, 2008, p. 14, 30-31). Like most performance standards, this threshold is inherently arbitrary. However, because WWC plays a large role in assessing the quality of impact studies in education, we adopted this threshold for assessing whether missing data methods yield bias that is large or small.

Specifically, in our simulations*, we classified an impact estimate as having "high bias" if the absolute value of the bias was greater than 0.05 standard deviations.* Because this threshold is based on the WWC's attrition standards, simulation results that show a particular method yields lower bias than 0.05 can be treated as evidence that the method produced estimates with a level of bias that is treated as acceptable by the WWC.[87]

B. Bias Standards for the Estimated Standard Errors of the Impact Estimates

Missing data can lead to biased standard errors, as well as biased impact estimates. In addition, impact and standard error estimates both contribute to the hypothesis test of whether an impact is statistically significant: the t-statistic equals the estimated impact divided by the estimated standard error. Therefore, for our simulations, we decided it would be useful to set standards for assessing the magnitude of the bias in the standard errors. In interpreting the results from the simulations, we apply these standards to assess whether a given missing data method produced a standard error for the impact estimate with "high bias" or "low bias."

Building on the chosen standards for impact estimates described in Section A, *we classified a standard error estimate as having "high bias" if it would yield a t-statistic with as much bias as the t-statistics that result from an impact estimate for which the*

[87] It is important to note that this does not mean that studies which employ the missing data method in question would necessary, if reviewed by the WWC, be determined to have met WWC's standards. In fact, there are no WWC standards for which missing data methods are acceptable and which methods are unacceptable (US ED, 2008).

absolute value of the bias is greater than 0.05 when the impact estimate has zero bias. In this way, we rely entirely on the WWC's attrition standard to determine whether to classify the bias in the impact estimate or standard error as large or small.

To calculate the bias thresholds for the estimated standard errors, let SE equal the true standard error of the impact estimate, given the extent of missing data and the choice of method for addressing missing data.[88] Then the t-statistic used to test the null hypothesis of zero impact is given in equation (1) below:

$$(1) \qquad t = \frac{0.20}{SE}$$

How much bias in the t-statistic is introduced by a bias in the impact estimate of 0.05 standard deviations? If the bias is positive—that is, the impact estimate converges to 0.25 instead of the true impact of 0.20—then equation (2) shows the value of the t-statistic that results from this bias. This equation shows that a positive bias of 0.05 standard deviations yields a t-statistic that is 25 percent larger than it would be with an unbiased estimate of the standard error:

$$(2) \qquad t^{+0.05} = \frac{(0.20 + 0.05)}{SE} = \frac{0.25}{SE} = \frac{1.25 \times 0.20}{SE} = 1.25 \times t$$

If the bias is negative—that is, the impact estimate converges to 0.15 instead of the true impact of 0.20—then equation (3) shows the value of the t-statistic that results from this bias. This equation shows that a negative bias of 0.05 standard deviations yields a t-statistic that is 25 percent smaller than it would be with an unbiased estimate of the standard error:

$$(3) \qquad t^{-0.05} = \frac{(0.20 - 0.05)}{SE} = \frac{0.15}{SE} = \frac{0.75 \times 0.20}{SE} = 0.75 \times t$$

The results from equations (2) and (3) can be used to set thresholds for bias in the standard errors. Equation (4) shows the magnitude of the standard error ($SE^{+0.05}$) necessary when there is no bias in the impact estimate to generate the same bias in the t-statistic as a positive 0.05 standard deviation bias in the impact estimate when there is no bias in the standard error:

$$(4) \qquad \frac{0.25}{SE} = \frac{0.20}{SE^{+0.05}} \text{ , which is equivalent to } \frac{SE^{+0.05}}{SE} = \frac{0.20}{0.25} = \frac{4}{5} = 0.80 .$$

This implies that the standard error would need to be 20 percent smaller than the true standard error to have the same effect on the t-statistic as a positive bias in the impact estimate of 0.05 standard deviations.

[88] Note that this is not the same as the standard error of the impact estimate that researchers would have obtained with complete data.

Similarly, equation (5) shows the magnitude of the standard error ($SE^{-0.05}$) necessary when there is no bias in the impact estimate to generate the same bias in the t-statistic as a negative 0.05 standard deviation bias in the impact estimate when there is no bias in the standard error:

$$(5) \qquad \frac{0.15}{SE} = \frac{0.20}{SE^{-0.05}} \text{ , which is equivalent to } \frac{SE^{-0.05}}{SE} = \frac{0.20}{0.15} = \frac{4}{3} = 1.33\,.$$

This implies that the standard error would need to be 33 percent larger than the true standard error to have the same effect on the t-statistic as a negative bias in the impact estimate of 0.05 standard deviations.