2023

AN EVALUATION OF METHODS FOR HANDLING MISSING DATA IN RANDOMIZED CONTROLLED TRIALS WITH OMITTED MODERATION EFFECTS

Elizabeth Pauley
University of Rhode Island, epauley@uri.edu

AN EVALUATION OF METHODS FOR HANDLING MISSING DATA IN RANDOMIZED CONTROLLED TRIALS WITH OMITTED MODERATION EFFECTS

BY

ELIZABETH M. PAULEY

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE

REQUIREMENTS FOR THE DEGREE OF

MASTER OF ARTS

IN

PSYCHOLOGY

UNIVERSITY OF RHODE ISLAND

2023

MASTER OF ARTS THESIS

OF

ELIZABETH M. PAULEY

APPROVED:

Thesis Committee:

Major Professor Manshu Yang

Nichea Spillane

Jing Wu

Brenton DeBoef
DEAN OF THE GRADUATE SCHOOL

UNIVERSITY OF RHODE ISLAND
2023

ABSTRACT

Randomized Control Trials (RCTs) are widely used in behavioral and health-related studies to evaluate the effectiveness of intervention strategies; however, missing data in RCTs are almost inevitable. In many RCT studies, the key focus is to examine the average treatment effect (ATE) within an entire population. Heterogenous treatment effects, often reflected in moderation effects of baseline personal attributes, do not typically get included in analyses. To handle missing data in RCTs, multiple imputation (MI) or inverse probability weighting (IPW) could be used. MI, although often preferred over IPW, may lead to biased ATE results when the probability of missingness depends on a moderator and the moderation effect is omitted from the imputation process. In contrast, IPW may produce imprecise results when the sample size is small. This study aims to evaluate the performance of MI via joint modeling (MI-JM), MI via chained equations (MI-CE), and IPW in estimating the ATE in RCTs with missing data and omitted moderation effects. A Monte Carlo simulation study is conducted to compare methods under various scenarios. Findings suggest that the use of MI-CE would be recommended across all study conditions with the presence of incomplete outcomes but fully observed covariates. IPW could be utilized with relatively large sample sizes and relatively a small number of covariates. Listwise deletion and MI-JM are not recommended for use in RCTs with missing data and omitted moderation effects.

ACKNOWLEDGMENTS

I am deeply grateful to my thesis advisor, Dr. Manshu Yang, for her invaluable guidance, support, and expert insights throughout the entire process of researching and writing this thesis. I would like to express my sincere gratitude to the members of my thesis committee, Dr. Nichea Spillane, Dr. Jing Wu, and Dr. Annemarie Vaccaro, for their thoughtful feedback and contributions. I also extend my appreciation to the staff and resources at The University of Rhode Island.

I am thankful to my friends and family for their encouragement, patience, and belief in me. Your constant support provided the motivation needed to overcome challenges and reach this milestone. This research would not have been possible without the contributions and support of all those mentioned above, and for that, I am truly grateful.

TABLE OF CONTENTS

ABSTRACT ii

ACKNOWLEDGMENTS iii

TABLE OF CONTENTS iv

LIST OF TABLES v

CHAPTER 1 1

INTRODUCTION 1

CHAPTER 2 10

METHODOLOGY 10

CHAPTER 3 15

FINDINGS 15

CHAPTER 4 19

DISCUSSION 19

LIST OF TABLES

Table 1. Relative Bias of Average Treatment Effect Estimates with 10% Missing Data, ATE=0.5, Small Moderation Effects ((β_{31} to β_{35} = 0.1), by Number of Omitted Moderators, Sample Size, and Method. 22

Table 2. Relative Bias of Average Treatment Effect Estimates with 10% Missing Data, ATE=0.5, Medium Moderation Effects ((β_{31} to β_{35} = 0.3), by Number of Omitted Moderators, Sample Size, and Method 23

Table 3. Relative Bias of Average Treatment Effect Estimates with 10% Missing Data, ATE=0.5, Large Moderation Effects ((β_{31} to β_{35} = 0.5), by Number of Omitted Moderators, Sample Size, and Method 24

Table 4. Relative Bias of Average Treatment Effect Estimates with 30% Missing Data, ATE=0.5, Small Moderation Effects ((β_{31} to β_{35} = 0.1), by Number of Omitted Moderators, Sample Size, and Method 25

Table 5. Relative Bias of Average Treatment Effect Estimates with 30% Missing Data, ATE=0.5, Medium Moderation Effects (β_{31} to β_{35} = 0.3), by Number of Omitted Moderators, Sample Size, and Method 26

Table 6. Relative Bias of Average Treatment Effect Estimates with 30% Missing Data, ATE=0.5, Large Moderation Effects (β_{31} to β_{35} = 0.5), by Number of Omitted Moderators, Sample Size, and Method 27

Table 7. Type 1 Error Rate of Detecting Average Treatment Effect with 10% Missing Data, ATE =0, Small Moderate Effects (β_{31} to β_{35} = 0.1), by Number of Omitted Moderators, Sample Size and Method 28

Table 8. Type 1 Error Rate of Detecting Average Treatment Effect with 10%

Missing Data, ATE =0, Medium Moderate Effects (β_{31} to β_{35} = 0.3), by Number of Omitted Moderators, Sample Size and Method.. 29

Table 9. Type 1 Error Rate of Detecting Average Treatment Effect with 10% Missing Data, ATE =0, Large Moderate Effects (β_{31} to β_{35} = 0.5), by Number of Omitted Moderators, Sample Size and Method.. 30

Table 10. Type 1 Error Rate of Detecting Average Treatment Effect with 30% Missing Data, ATE =0, Small Moderate Effects (β_{31} to β_{35} = 0.1), by Number of Omitted Moderators, Sample Size and Method.. 31

Table 11. Type 1 Error Rate of Detecting Average Treatment Effect with 30% Missing Data, ATE =0, Medium Moderate Effects (β_{31} to β_{35} = 0.3), by Number of Omitted Moderators, Sample Size and Method.. 32

Table 12. Type 1 Error Rate of Detecting Average Treatment Effect with 30% Missing Data, ATE =0, Large Moderate Effects (β_{31} to β_{35} = 0.5), by Number of Omitted Moderators, Sample Size and Method.. 33

CHAPTER 1

INTRODUCTION

Randomized Controlled Trials

Randomized Controlled Trials (RCTs) are commonly used in psychological studies to estimate the effectiveness of intervention strategies. For instance, Bjureberg et Al. (2022) utilized RCT to evaluate the effect of emotion regulation interventions on maladaptive anger, Bisby et. al. (2022) utilized a RCT to evaluate differences between therapist-guided and self-guided online treatments for anxiety and depression, and Brincks et al. (2022) utilized a RCT to evaluate a family-based intervention as a prevention method for adolescent alcohol use among Hispanic populations.

In a typical two-group pretest-posttest RCT, participants are randomly assigned either to a treatment group to receive the new treatment of interest or to a control group where they receive no treatment, placebo, or standard treatment. In many RCT applications, the primary goal is to estimate the average treatment effect (ATE) aggregated across an entire target population (Gerber & Green, 2012; Gomila & Clark, 2022; Holland, 1986), such as all adults in the United States with the opioid use disorder. The ATE is defined as the difference between the treatment and control groups in the mean treatment outcomes at posttest, while controlling for pretest outcome scores and other baseline covariates. Analysis of Covariance (ANCOVA) is a widely used method to estimate the ATE in RCTs (Howell, 2009; Maxwell, Delaney & Kelley, 2018). The ANCOVA model can be expressed as:

$$Y_i = \beta_0 + \beta_1 TREAT_i + \beta_{21} X_{1i} + \beta_{22} X_{2i} + \cdots + \beta_{2p} X_{pi} + \varepsilon_i \quad (1)$$

where Y_i represents the treatment outcome score of the i^{th} person measured at posttest, $TREAT_i$ is a binary indicator of the treatment assignment for the i^{th} person (1=treatment group and 0=control group), and $[X_{1i}, X_{2i}, \ldots, X_{pi}]$ is a set of covariates measured at pretest for the i^{th} person, such as the pretest outcome score, age, gender, race/ethnicity, and other personal characteristics. The regression coefficient β_1 represents the ATE and is the key parameter of interest, $[\beta_{21}, \beta_{22}, \ldots, \beta_{2p}]$ represent covariate effects on the posttest outcome, and ε_i is the error term. By including covariates highly predictive of the treatment outcome, the ANCOVA model allows researchers to detect the ATE more efficiently with greater statistical power (Maxwell, Delaney & Kelley, 2018).

Omitted Moderation Effects

In psychological studies, it is not uncommon that the treatment may work better for some individuals than others (Lee et. al., 2019; Marquardt et. al., 2022; Wachs et. al., 2022). For instance, less acculturated individuals may be less responsive to an intervention to reduce problematic drinking than those more acculturated. Such heterogeneous treatment effects, or in other words, moderation effects of treatment, could be captured by adding an interaction term — between the treatment indicator ($TREAT$) and the covariate as potential moderator — to the original ANCOVA model. If complete data are obtained for every person in an RCT, omitting such interaction effects in an ANCOVA model does not raise issues when estimating the ATE. However, when missing data are present and the probability of missingness is determined by the covariates that involve omitted moderation effects, the ATE estimates could be substantially biased if the missing data are not properly handled.

Missing Data Issues in Randomized Controlled Trials

Missing data are almost inevitable in RCTs and may occur for a variety of reasons. For example, participants may be unwilling to provide answers to some survey questions, miss an assessment session, or completely drop out of a study, thereby posing issues when analyzing collected incomplete data and drawing research conclusions (Schafer & Graham, 2002). To describe how missingness is potentially related to the data, Rubin (1976) classified the missing data mechanisms into three types, including missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR).

MCAR occurs when the probability of having a missing value does not depend on the primary research question or study variable values (e.g., participant falls ill during the study unrelated to any study measure). MAR occurs when the probability of missingness is not related to the unseen values of the incomplete variable itself but does depend on observed values of other measures in the study (e.g., individuals with higher observed scores of emotion dysregulation are more likely to skip questions in a personality scale). MNAR occurs when the probability of missingness depends on the unseen values of the incomplete variable itself or other unobserved variables related to the study variables (e.g., participants not filling out a substance use questionnaire due to current engagement with substances).

The current study will focus on analyzing MAR data where the probability of having a missing posttest outcome score depends on one or more measured covariates that potentially moderate the treatment effects, but the moderation effects (i.e., interaction terms) are not included in the primary analysis that examines the ATE.

Statistical Methods to Handle Missingness

Various methods of handling missing data can be utilized depending on the amount and type of missingness. For the purpose of this study, three methods will be evaluated, including inverse probability weighting (IPW), multiple imputation via joint modeling (MI-JM), and multiple imputation via chained equations (MI-CE).

Inverse Probability Weighting (IPW)

Until the development of more modern methods, psychologists often used ad hoc methods, such as complete-case analysis, to deal with missing data. An individual is considered as a “complete case” if their data on all the variables involved in the analysis (e.g., outcome and all the covariates in an ANCOVA model) are fully observed. Complete-case analysis is also referred to as the listwise deletion (LD) method because any records containing missing data are deleted entirely from the analysis. Complete-case analysis is the default method in most regression analysis software (White & Carlin, 2008), but it usually produces biased results unless the missingness mechanism is MCAR or the missingness depends entirely on covariates and the regression model is correctly specified (Little, Carpenter, & Lee, 2022; Johnson & Young, 2011; Schafer & Graham, 2002). To handle MAR data in RCTs, the IPW approach could be implemented, which still uses complete cases but gives more weight to some cases than others in the regression (e.g, ANCOVA) model (Little et al., 2022).

IPW involves providing weight to cases based upon the probability of them being complete (Seaman & White, 2011). Specifically, individuals with a higher probability of being a complete case will receive less weight than those with a lower probability of being complete. In Step 1 of the IPW approach, the probability of the i^{th} person being a complete case, namely P_i, is calculated using a logistic regression model:

$$P_i = \frac{\exp(\gamma_0 + \gamma_1 TREAT_i + \boldsymbol{\gamma_2 X_i} + \boldsymbol{\gamma_3} TREAT_i \boldsymbol{X_i})}{1 + \exp(\gamma_0 + \gamma_1 TREAT_i + \boldsymbol{\gamma_2 X_i} + \boldsymbol{\gamma_3} TREAT_i \boldsymbol{X_i})} \quad (2)$$

where $\boldsymbol{X_i} = [X_{1i}, X_{2i}, \ldots, X_{pi}]$ is a column vector of covariates measured at pretest for the i^{th} person, and $TREAT_i \boldsymbol{X_i} = [TREAT_i X_{1i}, TREAT_i X_{2i}, \ldots, TREAT_i X_{pi}]$ is a column vector of interaction terms between $TREAT$ and each covariate. In addition, $\gamma_{\mathbf{2}} = [\gamma_{21}, \gamma_{22}, \ldots, \gamma_{2p}]$ is a row vector of regression coefficients for $X_{\boldsymbol{i}}$ and $\gamma_{\mathbf{3}} = [\gamma_{31,} \gamma_{32}, \ldots \gamma_{3p}]$ is a row vector of regression coefficients for the interaction terms between $TREAT$ and each covariate. If the logistic regression model is mis-specified (e.g., the interaction term $TREAT_i \boldsymbol{X_i}$ is omitted), IPW may yield biased results. The weight for person i is then computed using the inverse probability formula:

$$w_i = \frac{1}{P_i} \quad (3)$$

In Step 2 of the IPW approach, a weighted regression model is fitted with only complete cases. The regression model uses the same specification as the ANCOVA model shown in Equation 1, except that each complete case is given the weight calculated in Step 1 of the IPW approach (Gomila & Clark, 2020).

Multiple Imputation (MI)

The MI approach has been increasingly used to deal with missing data in psychological studies in the past two decades (Schafer & Graham, 2002). MI is separated into two phases, namely an *imputation phase* and an *analysis and pooling phase* (Enders, 2010). The imputation phase involves using the distribution of the observed data to simulate multiple plausible values for each missing value, resulting in *K* versions of a complete dataset that replace (i.e., impute) missing values with simulated plausible values. In the analysis and pooling phase, each imputed data set is analyzed using the same complete-data inference model, and results from the *K* data sets are then combined via Rubin's Rules (Rubin, 1987) to obtain the overall estimates and standard errors of the parameters (e.g., ATE), which represent both the sample variation and the uncertainty surrounding missingness. Compared to complete-case analysis, MI has been found to produce less biased parameter estimates when data are MAR (Enders, 2010; Lieberman-Betz, et. al, 2014; Schafer & Graham, 2002). In addition, MI often yields more precise estimates than IPW, because the former includes all individuals (even those with partially missing data) whereas the latter only includes complete cases in the analyses.

However, as discussed in Seaman & White (2013), MI may yield biased results if the imputation model is mis-specified. The current study will examine the performance of two types of MI, namely joint modeling and chained equations, given a potential source of imputation model misspecification commonly encountered in RCTs, that is, when moderation effects are omitted from the inferential analysis model.

Multiple Imputation via Joint Modeling (MI-JM). MI-JM assumes all the variables included in the missing data imputation model follow a common joint distribution. In the context of RCT and ANCOVA, it is assumed that all the covariates and outcome are jointly normally distributed within the treatment (or control) group. Consequently, a multivariate normal distribution is used to predict/impute missing values in all incomplete variables simultaneously, based on the observed values of all variables. To illustrate, consider an RCT with an incomplete outcome Y and a completely observed covariate X, the joint imputation model can be expressed as

$$\binom{Y_{i,mis}}{X_i} \sim MVN\left(\begin{matrix}\alpha_{10,y} + \alpha_{11,y}TREAT_i \\ \alpha_{10,x} + \alpha_{11,x}TREAT_i\end{matrix}, \Sigma_e\right) \quad (4)$$

where $Y_{i,mis}$ represents the missing value of Y for the i^{th} person and Σ_e is the covariance matrix between Y and X. Of note, although this study focuses on normally distributed variables, the MI-JM approach can be readily extended to accommodate categorical variables by assuming an underlying normally distributed latent trait for discrete responses. As shown in Equation (4), in the joint modeling imputation model, while the means of X and Y are allowed to vary between the treatment and control groups, the covariance structure of X

and Y (i.e., Σ_e) are assumed equal between the two groups. In other words, the relationship between X and Y is assumed equal between groups, and nonlinear terms such as moderation effects (where X moderates the effect of treatment on Y) cannot be included in the imputation. Therefore, omitting moderation effects would result in mis-specified imputation model and potentially biased treatment effect estimates (Little et al., 2022).

Multiple Imputation via Chained Equations (MI-CE). MI-CE allows for more flexibility as compared to MI-JM due to its ability to include moderation effects or interaction terms into the imputation model. A key distinction between MI-JM and MI-CE is that the former draws replacement values for all incomplete variables from a common multivariate distribution, whereas the latter cycles through incomplete variables one at a time, drawing replacement values from a series of univariate conditional distributions. At each iteration step of the imputation process, missing values for a particular variable are filled in by drawing plausible values from a univariate conditional distribution, based on a regression model that uses the incomplete variable as outcome and all other variables as predictors (including filled-in values of predictor variables from a previous step). To illustrate, consider an RCT with an incomplete outcome Y and a completely observed covariate X. If the probability of missing an outcome value depends on the covariate and there is an interaction effect between the covariate and the treatment assignment, then the imputation model can be expressed as

$$Y_{i,miss} \sim N(\alpha_{20} + \alpha_{21}TREAT_i + \alpha_{22}X_i + \alpha_{23}TREAT_iX_i, \sigma_e^2). \quad (5)$$

Despite the increased flexibility, under certain conditions (i.e., small sample sizes, large number of interaction terms), MI-CE may produce imprecise results with large standard errors (Little et al., 2022).

Study Aims

There has been a dearth of investigation focusing on the performance of MI and IPW in RCTs with omitted moderation effects as well as the key factors influencing their performance. Consequently, the current study aimed to compare the performance of three methods in estimating ATE in RCTs with MAR data and omitted moderation effects, given various sample sizes and analysis/imputation model complexity levels. The three methods include:

1. the IPW approach that includes all possible moderation effects when computing the probability of being complete,
2. the MI approach via joint modeling (MI-JM) which does not include moderation effects when imputing missing values, and
3. the MI approach via chained equations (MI-CE) which includes all possible moderation effects when imputing missing values.

The results of this study will provide support for which of the three proposed methods performs the best under which conditions. It is hypothesized that IPW or MI-CE will outperform MI-JM, when the sample size is relatively large and the moderation effects are strong.

CHAPTER 2

METHODOLOGY

A Monte Carlo simulation study was conducted to investigate the performance of IPW, MI-JM, and MI-CE methods in estimating the average treatment effect in RCTs with MAR data and omitted moderation effects. In addition, to facilitate the comparisons, analyses based on the full complete dataset before generating any missing values (i.e., the Complete method) and the listwise deletion (LD) method were used to demonstrate the “best case scenario” and the “worst case scenario”, respectively.

Monte Carlo simulations are commonly used to investigate the magnitude of bias in parameter estimates (e.g., whether a statistical method consistently under- or over-estimate the population value of treatment effect), to determine whether a method maintains the Type I error rate at the desired level, and to examine statistical power. These evaluation goals cannot be achieved in empirical data analyses, where parameters (e.g., treatment effects) are estimated using real-world data collected from participants and the true values of parameters are unknown. Consequently, researchers cannot tell how close the estimated treatment effect from a sample is to the actual population value of treatment effect. In contrast, in simulation studies, parameters are estimated using a large number of simulated data sets where the true values of the population parameters are known, and hence researchers can determine whether and to what extent sample results are consistently below or above the population values.

Simulation Study Design

<u>*Complete Data Generation*</u>

Data was generated to simulate a RCT with two groups (a treatment group and a control group) and five covariates. Under each simulated condition, complete datasets with no missing data were generated using the following model:

$$Y_i = \beta_0 + \beta_1 TREAT_i + \beta_{21}X_{1i} + \beta_{22}X_{2i} + \beta_{23}X_{3i} + \beta_{24}X_{4i} + \beta_{25}X_{5i} + \beta_{31}TREAT_iX_{1i} + \beta_{32}TREAT_iX_{2i} + \beta_{33}TREAT_iX_{3i} + \beta_{34}TREAT_iX_{4i} + \beta_{35}TREAT_iX_{5i} + \varepsilon_i \quad (6)$$

where the ATE, or β_1, was fixed at either 0 or 0.5 to represent no treatment effect or a medium-sized treatment effect that is commonly found in psychological studies. Without loss of generality, the variances of the error term ε and the five covariates X_1 to X_5 were set to 1; the intercept of Y (β_0)was set to 3; the means of covariates X_1 to X_5 in both groups were set to 0. To ensure that bias would result if the moderation effect pertaining to a covariate is omitted from the analysis model, a medium sized correlation was set between each covariate and the outcome (Collins et al., 2001), with β_{21} to β_{25} set to 0.4. To investigate how the number of omitted moderation effects impact the analysis results, three conditions, with 1, 3, or 5 moderation effects, were examined. Under the condition with one moderation effect, only the first interaction term ($\beta_{31}TREAT_iX_{1i}$) in Equation 6 was included to generate data and $\beta_{32}TREAT_iX_{2i}$ to $\beta_{35}TREAT_iX_{5i}$ were removed from the data generation model. Similarly, under the conditions with 3 and 5 moderation effects, the first three interaction terms and all five interaction terms in Equation 6, respectively, were included. Without loss of generality, treatment and control groups were equally sized. For each person, the covariate values (X_{1i} to X_{5i}) were generated first, followed by the generation of the residual term (ε_i). Lastly, the corresponding outcome value (Y_i) was computed based on Equation 6. Data was generated using the statistical software R (R Core Team, 2021).

Missing Data Generation

Missing data was generated assuming MAR. The probability of missing the posttest outcome Y for the i^{th} person was dependent on all the five covariates and can be expressed as

$$Pr(R_i = 1) = \Phi(\eta_0 + \eta_1X_{1i}^* + \eta_2X_{2i}^* + \eta_3X_{3i}^* + \eta_4X_{4i}^* + \eta_5X_{5i}^* \quad (7)$$

where Φ represents the probit function, X_{1i}^* to X_{1i}^* are standardized scores of X_1 to X_5 (so that the variable representing the sum of X_1 to X_5 has a variance of 1), and R_i is a binary indicator with a value of 1 indicating Y value was missing and 0 indicating Y was observed. The coefficients η_1 to η_5 were set at 1.815 to allow a 0.5 correlation between the covariates and the probability of missing the outcome; η_0 was set at -3.22 or -1.31 to generate 10% or 30% of missing data in the outcome Y. A missing data rate of 10% to 30% was commonly seen in psychological studies, as reported in previous literature (e.g., Little et al., 2014). To determine if a person has missing Y value, the probability of missing Y for this person was first specified using Equation 7, and the missing indicator R for this person was then generated from a binomial distribution with success rate equal to their probability of missing Y. For persons with $R_i = 1$, their Y values were set to missing in the final generated data set.

Influencing factors examined in the simulation study

Five factors were manipulated in the simulations to investigate their impact on the ATE estimation. The factors included (1) population value of ATE (β_1 = 0 or 0.5), (2) number of omitted moderation effects (1, 3, or 5), (3) magnitude of omitted moderation effects (β_{31} to β_{35} = 0.1, 0.3, or 0.5), (4) sample size (n = 50, 100, 200, or 400), and (5) proportion of missing data in the outcome p_{miss} = 10% or 30%. A total of 2 × 3 × 3 × 4 × 2 = 144 conditions were examined. For each condition, 1,000 datasets were generated.

Methods used to analyze data

After generating the data, the ANCOVA model in Equation 1 was fitted to each dataset to estimate the ATE using the three methods of interest: IPW, MI-JM, and MI-CE, as well as the two additional comparison methods: Complete and LD. Inferential analyses were conducted using the lm function in R, the probability of being complete (when implementing IPW) was calculated using the glm function in R, and the multiple imputation was conducted using the R jomo package (Quartagno, 2022) for MI-JM and the R mice package (Van Buuren & Groothuis-Oudshoorn, 2011) for MI-CE.

Evaluation Statistics

The ATE was the primary parameter of interest for evaluating the estimation and inference process. Estimation bias of ATE was evaluated for all methods across all the simulation conditions and Type I error rate was evaluated when ATE = 0. The absolute bias, defined as the deviation of the estimated ATE — averaged across the 1,000 simulated datasets — from its true population value, was computed when ATE was zero, whereas relative bias, defined as the ratio of absolute bias to the true value of ATE (expressed in a percent format), was examined given nonzero ATE. A relative bias greater than 10% is typically considered as not acceptable (Finch et al., 1997; Kaplan, 1988). The empirical Type I error rate is defined as the proportion of significant ATE estimates among the 1,000 simulated datasets when the true ATE is zero. Given a nominal alpha level of 0.05, a Type I error rate higher than 0.075 is often considered as inflated, and a rate lower than 0.025 considered as deflated.

CHAPTER 3

FINDINGS

The results from the simulation study are organized as follows. The estimation bias of ATE across the five methods are presented first, followed by the Type I error rates obtained from the zero treatment effect conditions.

Estimation bias for average treatment effect

Bias with 10% Missing Data and ATE=0.5

Tables 1 to 3 present the summary information of relative bias for the ATE estimates from the five methods when the true treatment effect was nonzero (i.e., 0.5) and the proportion of missing data was 10%, with the magnitude of omitted moderation effects equal to 0.1, 0.3, and 0.5, respectively. In each table, results are stratified by sample size and number of omitted moderators. Relative biases in regard to estimating the ATE were small or negligible and never exceeded the threshold of 10%, regardless of the method, the sample size, the number of omitted moderators, or the magnitude of moderation effects evaluated.

Bias with 30% Missing Data and ATE=0.5

Tables 4 to 6 present the summary information of relative bias for the ATE estimates from the five methods when the true treatment effect was nonzero (i.e., 0.5) and the proportion of missing data was 30%, with small, medium, and large moderation effects (i.e., the magnitude of omitted moderation effects equal to 0.1, 0.3, and 0.5), respectively. Results are stratified by sample size and number of omitted moderators. As shown in Table 4, given small moderation effects, relative biases in regard to estimating ATE were negligible and never exceeded the threshold of 10%, regardless of the method, sample size, or number of omitted moderators evaluated.

Given medium-sized moderation effects (see Table 5), relative biases in estimating the ATE were minimal when the number of omitted moderators was one. However, when the number of omitted moderators was 3 or 5, LD and MI-JM resulted in biased ATE estimates, with relative biases greater than 10% across all four sample sizes.

When the omitted moderation effects were large (see Table 6), relative biases in estimating the ATE were still negligible when there was only one omitted moderator. However, when the number of omitted moderators was 3 or 5, LD and MI-JM resulted in severely biased ATE estimates across all four sample sizes, with approximately 40% relative biases when using MI-JM. In addition, ATE estimates via IPW showed relative biases greater than 10% when the sample size was small (n=50) and many moderators (5) were omitted from the inferential analyses.

Overall, as the number of omitted moderators increased or as the magnitude of the moderation effects increased, the bias in estimating the ATE became more pronounced when using LD or MI-JM. As the sample size decreased, the bias in estimating the ATE became more pronounced when using IPW but remained similar when using LD or MI-JM. Across conditions, MI-CE performed relatively well, producing minimal biases.

Given that absolute biases in estimating the ATE were similar when ATE = 0 and when ATE = 0.5, the results from the former are omitted here.

Type I Error for Detecting Average Treatment Effect

Type I Error Rates with 10% Missing Data

Tables 7 to 9 show the empirical Type I error rates in detecting the ATE, across the five methods, when the proportion of missing data was 10% and the magnitude of the omitted moderation effects was small (0.1), medium (0.3), and large (0.5), respectively. In each table, results are stratified by sample size and number of omitted moderators. With only 10% of missing data, Type I error rates were close to the nominal level of 0.05 across varying sample sizes and numbers of omitted moderators. The only exception was that in the presence of small sample size (n=50) and a single omitted moderator that had strong moderation effect, MI-JM led to inflated Type I error rate (0.077, see Table 9).

Type I Error Rates with 30% Missing Data

Tables 10 to 12 show the empirical Type I error rates in detecting the ATE, across the five methods, when the proportion of missing data was 30% and the magnitude of the omitted moderation effects was small (0.1), medium (0.3), and large (0.5), respectively. In each table, results are stratified by sample size and number of omitted moderators. With small moderation effects, the Type I error rates remained close to the nominal level of 0.05 across various methods. With medium and large moderation effects, as sample size and number of omitted moderators increased, LD and MI-JM resulted in more inflated Type I error rates

(>0.075, see Table 11 and Table 12, respectively). With medium moderation effects, small sample size (n=50) and 1 or 3 omitted moderators, IPW resulted in inflated Type I error rates (0.077 and 0.075, respectively, see Table 11). With large moderation effects, small sample size (n=50) and 3 omitted moderators, IPW resulted in inflated Type I error rates (0.077, see Table 12).

CHAPTER 4

DISCUSSION

With regard to the bias in estimating the ATE, LD and MI-JM led to increased biases as the number of omitted moderators and the magnitude of moderation increased, IPW produced more pronounced biases with smaller sample sizes (i.e., n=50), and MI-CE consistently resulted in minimal biases. As sample size and number of omitted moderators and magnitude of moderation increased, LD and MI-JM led to more severely inflated Type I error rates.

In the presence of only 10% missing data, all methods showed minimal biases and Type I error rates close to the nominal level of 0.05. This is consistent with previous literature stating that MI provides negligible benefits as compared to LD, given less than 5% missing data (Schafer, 1999), and substantial bias likely occurs in analyses with more than 10% missingness (Dong & Pend, 2013).

As expected, LD led to substantial biases and inflated Type I error rates with 30% missing data. In this study, the treatment effect was set to vary depending on the observed scores of the moderator(s), and the probability of missing the outcome Y was higher for individuals with higher scores of the moderator(s). Thus forth, the complete cases tend to be individuals with lower scores of the moderator(s), and the ATE was biased towards the treatment effect given lower scores of the moderator(s).

Similarly, MI-JM resulted in substantial biases and inflated Type I error rates. Given that MI-JM assumes multivariate normality and equal relationship between the covariates (e.g., potential moderators) and the outcome for the treatment and control groups, it was unable to incorporate moderation effect(s) into the imputation model, leading to biases and inflated Type I error rates. Of note, when the sample size was small (n=50), the Type I error rate resulting from the use of MI-JM was even higher than that from the use of LD. Despite multiple imputation typically being recommended over LD, if the imputation model is mis-specified, MI could be potentially more problematic than LD.

As the number of omitted moderators increased and moderation effects strengthened, the pitfalls of using LD and MI-JM became more salient, in regards to both bias and Type I error. When using LD or MI-JM, the magnitude of biases remained similar as the sample size increased; however, the Type I error rate became more severely inflated with larger sample sizes.

On the other hand, by giving more weight to cases that had a higher chance of being incomplete, IPW corrected the biases resulting from LD. Additionally, MI-CE exhibited minimal biases because it imputed the missing outcome values based on a correctly specified imputation model that included the omitted moderation effect(s).

The study was not without limitations. By only considering missing data in the outcome and assuming fully observed covariates and moderators, the performance of missing data handling methods given incomplete covariates was not examined. In addition, the impact of having a large number of covariates or moderators (e.g., more than 10) on the performance of MI and IPW was not investigated. Future research should aim to evaluate the performance of missing data handling methods when missing values are present in the covariates. Specifically, when covariates pertaining to the omitted moderation effects are partially missing, the MI-CE method may not perform as well as in the current study (Enders, Mistler & Keller, 2016; Enders, Hayes, & Du, 2018). As discussed in previous literature (Enders, Du & Keller, 2020; Lüdtke, Robitzsch, & West, 2020), with incomplete covariates that involve nonlinear effects, a substantive-model-compatible (SMC) imputation approach would be needed. The exploration of various MI approaches, including SMC imputations, could provide a more comprehensive comparison and useful guidance on which missing data handling method(s) should be used in RCTs with omitted moderation effects. Additionally, the performance of the IPW and MI methods in the context of small sample sizes (e.g., n=50) and a large number of covariates or omitted moderators (e.g., more than 10 covariates or moderators) warrants further investigation.

Table 1

Relative Bias of Average Treatment Effect Estimates with 10% Missing Data, ATE=0.5, Small Moderation Effects (β_{31} to β_{35} = 0.1), by Number of Omitted Moderators, Sample Size, and Method

Omitted Moderators	Method	*% Relative Bias*			
		n=50	*n=100*	*n=200*	*n=400*
1	Complete	0.60	0.97	-1.30	-0.18
	LD	0.02	0.80	-1.79	-0.46
	IPW	0.29	1.16	-1.47	-0.17
	MI-CE	0.51	0.87	-1.48	-0.18
	MI-JM	-0.01	0.82	-1.85	-0.46
3	Complete	-0.93	0.36	0.74	0.56
	LD	-1.63	-0.55	-0.10	-0.29
	IPW	-1.42	-0.16	0.88	0.65
	MI-CE	-0.78	0.27	0.91	0.60
	MI-JM	-1.61	-0.48	-0.13	-0.25
5	Complete	-0.45	-0.49	-0.24	-0.21
	LD	-2.17	-2.18	-1.90	-1.89
	IPW	-1.80	-0.96	-0.49	-0.40
	MI-CE	-0.89	-0.55	-0.25	-0.43
	MI-JM	-2.28	-2.14	-1.91	-1.91

Note: Complete = complete-data (pre-deletion) analysis; LD = listwise deletion; IPW = inverse probability weighting; MI-JM= multiple imputation via joint modeling; MI-CE= multiple imputation via chained equations; n= sample size; Relative Bias= the ratio of absolute bias to the true value in percentage format (relative bias >10% bolded).

Table 2

Relative Bias of Average Treatment Effect Estimates with 10% Missing Data, ATE=0.5, Medium Moderation Effects (β_{31} to β_{35} = 0.3), by Number of Omitted Moderators, Sample Sizeand Method

Omitted Moderators	Method	*% Relative Bias*			
		n=50	*n=100*	*n=200*	*n=400*
1	Complete	0.56	-1.68	0.50	-1.12
	LD	0.16	-2.29	-0.71	-2.17
	IPW	0.23	-1.69	0.28	-1.25
	MI-CE	1.02	-1.28	0.19	-1.23
	MI-JM	0.15	-2.32	-0.65	-2.23
3	Complete	-4.09	-0.32	0.59	0.39
	LD	-6.71	-3.26	-2.55	-2.51
	IPW	-6.05	-1.30	0.07	0.32
	MI-CE	-3.79	-0.39	0.32	0.31
	MI-JM	-6.42	-3.22	-2.58	-2.47
5	Complete	2.71	0.37	-1.57	1.10
	LD	-1.12	-4.12	-6.20	-3.53
	IPW	-0.20	-1.25	-1.90	1.03
	MI-CE	3.88	0.53	-1.57	1.11
	MI-JM	-1.06	-4.13	-6.22	-3.53

Note: Complete = complete-data (pre-deletion) analysis; LD = listwise deletion; IPW = inverse probability weighting; MI-JM= multiple imputation via joint modeling; MI-CE= multiple imputation via chained equations; n= sample size; Relative Bias= the ratio of absolute bias to the true value in percentage format (relative bias >10% bolded).

Table 3

Relative Bias of Average Treatment Effect Estimates with 10% Missing Data, ATE=0.5, Large Moderation Effects (β_{31} to β_{35} = 0.5), by Number of Omitted Moderators, Sample Size, and Method

Omitted Moderators	Method	*% Relative Bias*			
		n=50	*n=100*	*n=200*	*n=400*
1	Complete	-3.54	-0.44	-0.49	0.65
	LD	-5.23	-2.96	-1.95	-0.94
	IPW	-5.26	-1.80	-0.58	0.58
	MI-CE	-4.30	-1.57	-0.42	0.58
	MI-JM	-5.17	-2.97	-1.99	-0.88
3	Complete	0.41	-0.47	-0.25	0.02
	LD	-3.39	-4.95	-4.76	-4.87
	IPW	-2.50	-2.00	-0.58	-0.26
	MI-CE	1.16	-0.19	-0.28	-0.19
	MI-JM	-3.36	-4.94	-4.76	-4.88
5	Complete	0.12	0.35	2.19	0.86
	LD	-6.68	-6.79	-5.65	-7.16
	IPW	-5.28	-1.82	1.55	0.68
	MI-CE	0.40	0.87	2.33	0.68
	MI-JM	-6.69	-6.82	-5.72	-7.14

Note: Complete = complete-data (pre-deletion) analysis; LD = listwise deletion; IPW = inverse probability weighting; MI-JM= multiple imputation via joint modeling; MI-CE= multiple imputation via chained equations; n= sample size; Relative Bias= the ratio of absolute bias to the true value in percentage format (relative bias >10% bolded).

Table 4

Relative Bias of Average Treatment Effect Estimates with 30% Missing Data, ATE=0.5, Small Moderation Effects (β_{31} to β_{35} = 0.1), by Number of Omitted Moderators, Sample Size, and Method

Omitted Moderators	Method	*% Relative Bias*			
		n=50	*n=100*	*n=200*	*n=400*
1	Complete	-2.82	1.92	0.14	0.83
	LD	-4.32	1.56	-1.28	-1.20
	IPW	-3.14	3.22	0.04	0.47
	MI-CE	-2.78	2.85	0.03	0.47
	MI-JM	-4.49	1.58	-1.12	-1.17
3	Complete	3.44	0.57	0.11	0.69
	LD	-1.90	-3.40	-5.46	-4.10
	IPW	0.81	0.88	-1.04	0.43
	MI-CE	3.46	1.49	-0.91	0.46
	MI-JM	-2.32	-3.43	-5.35	-4.25
5	Complete	1.74	0.08	0.05	0.06
	LD	-6.52	-7.89	-7.91	-7.81
	IPW	-1.99	-0.44	-0.23	-0.04
	MI-CE	0.96	-0.03	0.04	0.07
	MI-JM	-6.28	-7.70	-7.92	-7.78

Note: Complete = complete-data (pre-deletion) analysis; LD = listwise deletion; IPW = inverse probability weighting; MI-JM= multiple imputation via joint modeling; MI-CE= multiple imputation via chained equations; n= sample size; Relative Bias= the ratio of absolute bias to the true value in percentage format (relative bias >10% bolded).

Table 5

Relative Bias of Average Treatment Effect Estimates with 30% Missing Data, ATE=0.5, Medium Moderation Effects (β_{31} to β_{35} = 0.3), by Number of Omitted Moderators, Sample Size, and Method

Omitted Moderators	Method	*% Relative Bias* n=50	*n=100*	*n=200*	*n=400*
1	Complete	-1.44	0.67	-0.80	-0.29
	LD	-7.33	-3.27	-4.58	-4.99
	IPW	-3.78	1.24	-0.10	-0.31
	MI-CE	-2.95	0.95	-0.17	-0.34
	MI-JM	-7.28	-3.32	-4.49	-5.08
3	Complete	0.19	2.85	-0.54	0.43
	LD	**-15.59**	**-10.49**	**-14.68**	**-13.42**
	IPW	-6.52	2.89	-0.80	0.57
	MI-CE	-0.62	3.82	-0.75	0.57
	MI-JM	**-15.79**	**-10.49**	**-14.67**	**-13.35**
5	Complete	6.37	-0.61	-1.89	-0.27
	LD	**-19.53**	**-23.71**	**-24.68**	**-22.86**
	IPW	-3.38	-1.89	-1.95	0.12
	MI-CE	5.15	0.77	-1.22	0.26
	MI-JM	**-19.57**	**-23.49**	**-24.60**	**-22.82**

Note: Complete = complete-data (pre-deletion) analysis; LD = listwise deletion; IPW = inverse probability weighting; MI-JM= multiple imputation via joint modeling; MI-CE= multiple imputation via chained equations; n= sample size; Relative Bias= the ratio of absolute bias to the true value in percentage format (relative bias >10% bolded)

Table 6

Relative Bias of Average Treatment Effect Estimates with 30% Missing Data, ATE=0.5, Large Moderation Effects ($\beta_{31}\ to\ \beta_{35}$ = 0.5), by Number of Omitted Moderators, Sample Size, and Method

Omitted Moderators	Method	*% Relative Bias*			
		n=50	*n=100*	*n=200*	*n=400*
1	Complete	1.65	1.80	-0.49	-1.39
	LD	-6.27	-5.83	-8.31	-9.59
	IPW	-1.69	0.95	-1.09	-1.95
	MI-CE	2.26	2.07	-0.64	-1.77
	MI-JM	-6.22	-5.57	-7.98	-9.60
3	Complete	-1.52	3.53	0.07	-0.61
	LD	**-23.88**	**-19.87**	**-22.93**	**-23.66**
	IPW	-9.51	1.71	-0.64	-0.63
	MI-CE	-1.79	3.69	0.00	-0.56
	MI-JM	**-23.44**	**-19.73**	**-22.97**	**-23.71**
5	Complete	2.68	-1.24	-1.78	-0.13
	LD	**-38.39**	**-40.09**	**-39.83**	**-39.11**
	IPW	**-11.87**	-4.35	-2.31	-0.19
	MI-CE	3.53	-1.26	-1.31	0.16
	MI-JM	**-37.54**	**-40.15**	**-39.93**	**-39.09**

Note: Complete = complete-data (pre-deletion) analysis; LD = listwise deletion; IPW = inverse probability weighting; MI-JM= multiple imputation via joint modeling; MI-CE= multiple imputation via chained equations; n= sample size; Relative Bias= the ratio of absolute bias to the true value in percentage format (relative bias >10% bolded).

Table 7

Type 1 Error Rate of Detecting Average Treatment Effect with 10% Missing Data, ATE=0, Small Moderation Effects (β_{31} to β_{35} = 0.1), by Number of Omitted Moderators, Sample Size and Method

Omitted Moderators	Method	*Type 1 Error Rate*			
		n=50	*n=100*	*n=200*	*n=400*
1	Complete	0.062	0.061	0.048	0.049
	LD	0.056	0.056	0.047	0.049
	IPW	0.057	0.061	0.042	0.048
	MI-CE	0.056	0.053	0.044	0.049
	MI-JM	0.062	0.058	0.049	0.051
3	Complete	0.052	0.057	0.047	0.052
	LD	0.049	0.053	0.048	0.058
	IPW	0.050	0.054	0.047	0.060
	MI-CE	0.048	0.052	0.044	0.057
	MI-JM	0.053	0.059	0.049	0.058
5	Complete	0.051	0.054	0.037	0.056
	LD	0.051	0.050	0.042	0.049
	IPW	0.051	0.053	0.041	0.051
	MI-CE	0.048	0.053	0.042	0.047
	MI-JM	0.065	0.054	0.047	0.050

Note: Complete = complete-data (pre-deletion) analysis; LD = listwise deletion; IPW = inverse probability weighting; MI-JM= multiple imputation via joint modeling; MI-CE= multiple imputation via chained equations; n= sample size; Type 1 Error rates higher than 0.075 are presented in bold and Type 1 error rates lower than 0.025 are presented in italic and bold.

Table 8

Type 1 Error of Average Treatment Effect Estimates with 10% Missing Data, ATE=0, Medium Moderation Effects (β_{31} to β_{35} = 0.3), by Number of Omitted Moderators, Sample Size, and Method

Omitted Moderators	Method	*Type 1 Error Rate*			
		n=50	*n=100*	*n=200*	*n=400*
1	Complete	0.046	0.055	0.062	0.052
	LD	0.051	0.054	0.066	0.049
	IPW	0.052	0.054	0.066	0.049
	MI-CE	0.047	0.055	0.064	0.051
	MI-JM	0.058	0.058	0.066	0.050
3	Complete	0.056	0.044	0.057	0.047
	LD	0.053	0.051	0.056	0.046
	IPW	0.052	0.053	0.062	0.046
	MI-CE	0.048	0.047	0.058	0.049
	MI-JM	0.061	0.052	0.060	0.051
5	Complete	0.046	0.057	0.049	0.051
	LD	0.052	0.058	0.055	0.055
	IPW	0.049	0.061	0.051	0.053
	MI-CE	0.045	0.057	0.049	0.054
	MI-JM	0.061	0.062	0.055	0.055

Note: Complete = complete-data (pre-deletion) analysis; LD = listwise deletion; IPW = inverse probability weighting; MI-JM= multiple imputation via joint modeling; MI-CE= multiple imputation via chained equations; n= sample size; Type 1 Error rates higher than 0.075 are presented in bold and Type 1 error rates lower than 0.025 are presented in italic and bold.

Table 9

Type 1 Error of Average Treatment Effect Estimates with 10% Missing Data, ATE=0, Large Moderation Effects (β_{31} to β_{35} = 0.5), by Number of Omitted Moderators, Sample Size, and Method

Omitted Moderators	Method	*Type 1 Error Rate* *n=50*	*n=100*	*n=200*	*n=400*
1	Complete	0.061	0.059	0.044	0.044
	LD	0.067	0.061	0.048	0.044
	IPW	0.066	0.065	0.047	0.044
	MI-CE	0.068	0.059	0.043	0.046
	MI-JM	0.077	0.069	0.050	0.042
3	Complete	0.048	0.051	0.049	0.045
	LD	0.047	0.050	0.050	0.048
	IPW	0.045	0.047	0.042	0.045
	MI-CE	0.047	0.050	0.043	0.042
	MI-JM	0.059	0.061	0.051	0.051
5	Complete	0.054	0.057	0.042	0.045
	LD	0.053	0.059	0.038	0.063
	IPW	0.052	0.053	0.041	0.047
	MI-CE	0.051	0.053	0.041	0.051
	MI-JM	0.057	0.061	0.041	0.062

Note: Complete = complete-data (pre-deletion) analysis; LD = listwise deletion; IPW = inverse probability weighting; MI-JM= multiple imputation via joint modeling; MI-CE= multiple imputation via chained equations; n= sample size; Type 1 Error rates higher than 0.075 are presented in bold and Type 1 error rates lower than 0.025 are presented in italic and bold.

Table 10

Type 1 Error of Average Treatment Effect Estimates with 30% Missing Data, ATE=0, Small Moderation Effects (β_{31} to β_{35} = 0.1), by Number of Omitted Moderators, Sample Size, and Method

Omitted Moderators	Method	*Type 1 Error Rate* *n=50*	*n=100*	*n=200*	*n=400*
1	Complete	0.053	0.050	0.046	0.066
	LD	0.053	0.051	0.054	0.051
	IPW	0.067	0.056	0.056	0.047
	MI-CE	0.032	0.043	0.048	0.052
	MI-JM	0.058	0.057	0.053	0.049
3	Complete	0.054	0.044	0.047	0.052
	LD	0.044	0.046	0.051	0.048
	IPW	0.049	0.059	0.056	0.057
	MI-CE	0.033	0.046	0.050	0.049
	MI-JM	0.056	0.046	0.054	0.050
5	Complete	0.053	0.057	0.053	0.049
	LD	0.048	0.058	0.057	0.063
	IPW	0.048	0.065	0.053	0.043
	MI-CE	0.038	0.054	0.047	0.043
	MI-JM	0.053	0.063	0.062	0.060

Note: Complete = complete-data (pre-deletion) analysis; LD = listwise deletion; IPW = inverse probability weighting; MI-JM= multiple imputation via joint modeling; MI-CE= multiple imputation via chained equations; n= sample size; Type 1 Error rates higher than 0.075 are presented in bold and Type 1 error rates lower than 0.025 are presented in italic and bold.

Table 11

Type 1 Error of Average Treatment Effect Estimates with 30% Missing Data, ATE=0, Medium Moderation Effects (β_{31} to β_{35} = 0.3), by Number of Omitted Moderators, Sample Size, and Method

Omitted Moderators	Method	*Type 1 Error Rate*			
		n=50	*n=100*	*n=200*	*n=400*
1	Complete	0.065	0.052	0.051	0.050
	LD	0.064	0.048	0.056	0.065
	IPW	**0.077**	0.058	0.051	0.056
	MI-CE	0.048	0.046	0.048	0.054
	MI-JM	0.070	0.051	0.055	0.068
3	Complete	0.063	0.052	0.049	0.047
	LD	0.062	0.047	**0.081**	**0.094**
	IPW	**0.075**	0.051	0.057	0.048
	MI-CE	0.039	0.041	0.059	0.043
	MI-JM	**0.077**	0.053	**0.081**	**0.089**
5	Complete	0.040	0.055	0.058	0.050
	LD	0.066	0.078	**0.103**	**0.155**
	IPW	0.061	0.054	0.053	0.042
	MI-CE	0.040	0.050	0.054	0.048
	MI-JM	0.068	**0.089**	**0.118**	**0.158**

Note: Complete = complete-data (pre-deletion) analysis; LD = listwise deletion; IPW = inverse probability weighting; MI-JM= multiple imputation via joint modeling; MI-CE= multiple imputation via chained equations; n= sample size; Type 1 Error rates higher than 0.075 are presented in bold and Type 1 error rates lower than 0.025 are presented in italic and bold.

Table 12

Type 1 Error of Average Treatment Effect Estimates with 30% Missing Data, ATE=0, Large Moderation Effects (β_{31} to β_{35} = 0.5), by Number of Omitted Moderators, Sample Size, and Method

Omitted Moderators	Method	*Type 1 Error Rate* *n=50*	*n=100*	*n=200*	*n=400*
1	Complete	0.044	0.043	0.056	0.054
	LD	0.043	0.051	0.054	0.062
	IPW	0.054	0.048	0.045	0.052
	MI-CE	0.036	0.045	0.041	0.050
	MI-JM	0.049	0.053	0.053	0.059
3	Complete	0.052	0.039	0.040	0.041
	LD	0.074	0.063	**0.080**	**0.148**
	IPW	**0.077**	0.051	0.039	0.037
	MI-CE	0.051	0.046	0.044	0.040
	MI-JM	**0.084**	0.067	**0.081**	**0.150**
5	Complete	0.038	0.041	0.059	0.042
	LD	0.057	**0.092**	**0.143**	**0.219**
	IPW	0.057	0.041	0.046	0.038
	MI-CE	0.037	0.037	0.046	0.048
	MI-JM	**0.075**	**0.094**	**0.147**	**0.220**

Note: Complete = complete-data (pre-deletion) analysis; LD = listwise deletion; IPW = inverse probability weighting; MI-JM= multiple imputation via joint modeling; MI-CE= multiple imputation via chained equations; n= sample size; Type 1 Error rates higher than 0.075 are presented in bold and Type 1 error rates lower than 0.025 are presented in italic and bold.

BIBLIOGRAPHY

Azur, M. J., Stuart, E. A., Frangakis, C., & Leaf, P. J. (2011). Multiple imputation by chained equations: What is it and how does it work? *International Journal of Methods in Psychiatric Research*, *20*(1), 40–49. https://doi.org/10.1002/mpr.329

Bisby, M. A., Scott, A. J., Hathway, T., Dudeney, J., Fisher, A., Gandy, M., Heriseanu, A. I., Karin, E., Titov, N., & Dear, B. F. (2022). Sudden gains in therapist-guided versus self-guided online treatments for anxiety or depression. *Journal of Consulting and Clinical Psychology*. *90*(11), 861-871. https://doi.org/10.1037/ccp0000771

Bjureberg, J., Ojala, O., Berg, A., Edvardsson, E., Kolbeinsson, Ö., Molander, O., Morin, E., Nordgren, L., Palme, K., Särnholm, J., Wedin, L., Rück, C., Gross, J. J., & Hesser, H. (2022). Targeting maladaptive anger with brief therapist-supported internet-delivered emotion regulation treatments: A randomized controlled trial. *Journal of Consulting and Clinical Psychology*. Advance online publication. https://doi.org/10.1037/ccp0000769

Brincks, A., Perrino, T., Estrada, Y., & Prado, G. (2022). Preventing alcohol use among Hispanic adolescents through a family-based intervention: The role of parent alcohol misuse. *Journal of Family Psychology*. 37(1), 105-109. https://doi.org/10.1037/fam0001038

Dong, & Peng, C.Y.J. (2013). Principled missing data methods for researchers. *SpringerPlus, 2*(1), 222-222. https://doi.org/10.118/2193 1801-2-222

Enders, C. K. (2010). *Applied missing data analysis*. New York, NY: Guilford.

Enders, C.K.. (2001). The performance of the full information maximum likelihood estimator in multiple regression models with missing data. *Educational and Psychological Measurement*, *61*(5), 713–740. https://doi.org/10.1177/00131644016150001Enders, C.K., & Bandalos, D. L. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural Equation Modeling*, *8*(3), 430–457. https://doi.org/10.1207/S15328007SEM0803_5

Enders, C.K., Du, H., & Keller, B.T. (2020). A model-based imputation procedure for multilevel regression models with random coefficients, interaction effects, and nonlinear terms. *Psychological Methods,* 25(1), 88-112.

Enders, C.K., Hayes, T., & Du, H. (2018). A comparison of multilevel imputation schemes for random coefficient models: Fully conditional specification and joint model imputation with random covariance matrices. *Mutlivariate Behavioral Research, 53(5),* 695-713.

Enders, C.K., Mistler, S.A., & Keller, B.T. (2016). Multilevel multiple imputation: A review and evaluation of joint modeling and chained equations imputation. *Psychological Methods, 21(2)*, 222-240.

Finch, J. F., West, S. G., & MacKinnon, D. P. (1997). Effects of sample size and nonnormality on the estimation of mediated effects in latent variable models. *Structural Equation Modeling: A Multidisciplinary Journal*, *4*(2), 87–107. https://doi.org/10.1080/10705519709540063

Gerber, A.S., & Green, D.P. (2012). *Field experiments: Design, analysis and interpretation*. New York, NY: Norton

Gomila, R., & Clark, C. S. (2022). Missing data in experiments: Challenges and solutions. Psychological Methods, 27(2), 143 155. https://doi.org/10.1037/met0000361

Graham, J. W., Olchowski, A. E., & Gilreath, T. D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science*, 8, 206–213. https://doi.org/10.1007/s11121-007-0070-9

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81*(396).945 960. https://doi.org/10.1080/01621459.1986.10478354.

Howell, D. C. (2009) *Statistical methods for psychology* (7th ed.). Belmont: Cengage Wadsworth.

Johnson, D. R., & Young, R. (2011). Toward best practices in analyzing datasets with missing data: Comparisons and recommendations. *Journal of Marriage and Family*, *73*(5), 926–945. https://doi.org/10.1111/j.1741-3737.2011.00861.x

Kaplan, D. (1988). The impact of specification error on the estimation, testing, and improvement of structural equation models. *Multivariate Behavioral Research*, *23* (1), 69–86. https://doi.org/10.1207/s15327906mbr2301_4

Lachin, J. M. (2000). Statistical considerations in the intent-to-treat principle. *Controlled Clinical Trials*, *21*(3), 167–189. https://doi.org/10.1016/S0197-2456(00)00046-5

Lee, C. S., Colby, S. M., Rohsenow, D. J., Martin, R., Rosales, R., McCallum, T. T., Falcon, L., Almeida, J., & Cortés, D. E. (2019). A randomized controlled trial of motivational interviewing tailored for heavy drinking latinxs. *Journal of Consulting and Clinical Psychology*, *87*(9), 815–830. https://doi.org/10.1037/ccp0000428

Lieberman-Betz, Yoder, P., Stone, W. L., Nahmias, A. S., Carter, A. S., Celimli-Aksoy, S., & Messinger, D. S. (2014). An illustration of using multiple imputation versus listwise deletion analyses: The effect of Hanen's "more than words" on parenting stress. *American Journal on Intellectual and Developmental Disabilities*, *119*(5), 472–491. https://doi.org/10.1352/1944-7558-119.5.472

Little T. D., Jorgensen, T. D., Lang, K. M., & Moore, E. W. G. (2014). On the joys of Missing data. Journal of Pediatric Psychology, 39(2), 151–162. https://doi.org/10.1093/jpepsy/jst048

Little, R.J., An, H., Johanns, J., & Giordani, B. (2000). A comparison of subset selection and analysis of covariance for the adjustment of confounders. *Psychological Methods, 5*(4), 459-476. https://doi.org/10.1037/1082-989X.5.4.459

Little, R.J., Carpenter, J. R., & Lee, K. J. (2022). A comparison of three popular methods for handling missing data: Complete-case analysis, inverse probability weighting, and multiple imputation. *Sociological Methods & Research*. https://doi.org/10.1177/00491241221113873

Lüdtke, O., Robitzsch, A., & West, S. G. (2020). Regression models involving nonlinear effects with missing data: A sequential modeling approach using Bayesian estimation. *Psychological Methods,* 25(2), 157-181

Madley-Dowd, Hughes, R., Tilling, K., & Heron, J. (2019). The proportion of Missing data should not be used to guide decision on multiple imputation. *Journal of Clinical Epidemiology, 110,* 63-73. https://doi.org/10.1016/j.jclinepi.2019.02.016

Marquardt, C. A., Chu, C., Hill, J. E., Venables, N. C., Kuzenski, L., Davenport, N. D., Disner, S. G., Finn, J. A., Gilmore, C. S., Erbes, C. R., & Urošević, S. (2022). Evaluating resilience in response to COVID-19 pandemic stressors among veteran mental health outpatients. *Journal of Psychopathology and Clinical Science*, *132*(1), 26-37. https://doi.org/10.1037/abn0000789

Maxwell, S. E., Delaney, H. D., & Kelley, K. (2018). *Designing experi-ments and analyzing data: A model comparison perspective* (3rd ed.). New York: Routledge.

Quartagno, M., Grund, S., & Carpenter, J. (2022). *Jomo: A package for Mult level Joint Modelling Multiple Imputation*. https://CRAN.R-project.org/package=jomo.

R Core Team (2021). R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria.* https://www.R-project.org/.

Rivera, R., Pérez, C.M., Martínez, M. N., & Suárez, E. (2017). Weighted least-squares linear regression. In *Applications of regression models in epidemiology* (pp. 117–128). John Wiley & Sons, Inc. https://doi.org/10.1002/9781119212515.ch8

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592. https://doi.org/10.1093/biomet/63.3.581

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. Hoboken, NJ: Wiley.

Savalei, & Rhemtulla, M. (2012). On obtaining estimates of the fraction of missing information from full information maximum likelihood. *Structural Equation Modeling*, *19*(3), 477–494. https://doi.org/10.1080/10705511.2012.687669

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

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*(2), 147-177. https://doi.org/10.1037/1082-989X.7.2.147

Schlomer, G. L., Bauman, S., & Card, N. A. (2010). Best practices for missing data management in counseling psychology. *Journal of Counseling Psychology*, *57*(1), 1–10. https://doi.org/10.1037/a0018082

Seaman, S. R., White, I. R., Copas, A. J., & Li, L. (2012). Combining multiple Imputation and inverse-probability weighting. *Biometrics, 68*(1), 129 137.. https://doi.org/10.1111/j.1541-0420.2011.01666.x.

Seaman, S.R., & White, I.R. (2013). Review of inverse probability weighting for dealing with missing data. *Stat Methods Med Res*, 278-95. doi: 10.1177/0962280210395740

Sinharay, S., Stern, H. S., & Russell, D. (2001). The use of multiple imputation for the analysis of missing data. *Psychological Methods*, *6*(4), 317–329. https://doi.org/10.1037/1082-989X.6.4.317

Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*(3), 1-67. https://doi.org/10.18637/jss.v045.i03

Wachs, S., Bilz, L., Wettstein, A., Wright, M. F., Kansok-Dusche, J., Krause, N., & Ballaschk, C. (2022). Associations between witnessing and perpetrating online hate speech among adolescents: Testing moderation effects of moral disengagement and empathy. *Psychology of Violence*, *12*(6), 371–381. https://doi.org/10.3390/ijerph16203992