THE IMPACT OF MULTIPLE IMPUTATION ON THE TYPE II ERROR RATE OF THE T TEST

by

TAMMY A. GRACE

DISSERTATION

Submitted to the Graduate School

of Wayne State University,

Detroit, Michigan

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

2016

MAJOR: EDUCATION EVALUATION AND RESEARCH

Approved By:

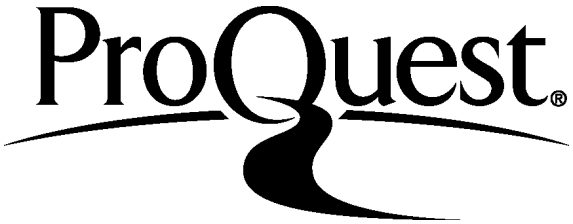
_____

Advisor Date

_____

_____

_____



DEDICATION

This work is dedicated, with love, to my Egg.

ACKNOWLEDGMENTS

I am indebted to my advisor, Dr. Shlomo Sawilowsky, for his invaluable support throughout my doctoral program. I feel extremely privileged to have an esteemed, Nobel Peace Prize nominee as a mentor.

I would also like to thank my original dissertation committee member, Dr. Barry Markman, and my replacement members, Dr. Claude Schochet and Associate Dean Thomas Edwards, who graciously stepped in when my previous members, Drs. Donald Marcotte and Gail Fahoome, passed away.

A debt of gratitude is due to Dr. Lawrence Radine and Dr. Fred Strale, not only for the academic support I received from them, but also the sincere friendship they have given me since the very beginning of my journey. A special thank you is also due to David Marso for sharing his programming expertise.

I also want to thank my husband and daughter for their endless support and unwavering encouragement.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1 INTRODUCTION

Missing data threaten the validity of clinical trials, yet receive little attention in the literature (O'Neill & Temple, 2012; Wood, White, & Thompson, 2004). Poor approaches for treating missing values can produce biased estimates, distorted statistical power, and invalid conclusions (Acock, 2005; Enders, 2010; Fichman & Cummings, 2003; Graham, 2012). These consequences are especially serious for phase III trials, which are intended to provide evidence of the efficacy and safety of medical treatments. The use of inadequate missing data methods can also impede the construction of valid prognostic models (Burton & Altman, 2004), undermine random assignment in randomized controlled trials (RCTs), and violate the intention to treat (ITT) approach to the analysis of clinical trial data.

Statistical power (rejecting a false null hypothesis) has also received less attention than it deserves (Cohen, 1962, 1988, 1990; Murphy, Myors, & Wolach, 2009). The literature is replete with meta-analyses demonstrating the shockingly low (e.g., .25) power of clinical trials (e.g., Button et al., 2013; Moher, Dulberg, & Wells, 1994; Tsang, Colley, & Lynd, 2009). The almost total lack of attention to Type II error (failure to reject a false null hypothesis) and its consequences have "worrying implications" (Williams & Seed, 1992, p. 321). In clinical research, a high probability of Type II error can lead to the underreporting of serious adverse events (e.g., death, major bleeding, serious infections) and erroneous conclusions of equivalent toxicity (Tsang et al., 2009). The conclusions derived from underpowered studies are often contradictory (Howard, Maxwell, & Fleming, 2000; Maxwell, 2004; Rossi, 1990) and make it difficult to draw coherent clinical inferences from the literature (Maxwell, 2004). Failing to detect the effects of treatments or interventions may also contribute to the premature termination of potentially valuable research (Cohen, 1962; Williams & Seed, 1993; Yuen & Pope, 2008) and

this is especially true when Type II errors are committed in exploratory studies involving innovative designs or small treatment effects (Chase & Chase, 1976; Freiman, Thomas, Chalmers, Smith, & Kuebler, 1978; Woods, Rippeth, Conover, Carey, Parsons, & Tröster, 2006).

Unfortunately, the consequences of underpowered studies are often exacerbated by the use of outdated missing data techniques. Listwise deletion (the practice of discarding cases with one or more missing values) has been shown to drastically increase the probability of a Type II error, yet still continues to dominate the RCT literature (Mackinnon, 2010; Wood et al., 2004), even in areas such as cancer research (Burton & Altman, 2004). Multiple appeals have called for the abandonment of traditional approaches (including listwise deletion) in favor of more principled methods (see Chapter 2).

Multiple imputation (MI) is a promising approach to treating missing data. First proposed by Rubin (1976) and elaborated in 1987, MI replaces missing values with $m > 1$ sets of imputed values, resulting in m complete datasets. Each of the datasets is analyzed and the results are combined to yield one set that reflects both within- and between-imputation uncertainty.

Initial MI procedures assumed a large joint model for variables (e.g., a joint normal distribution). As almost all datasets have mixtures of incomplete categorical and continuous variables, this assumption rarely holds in practice. Fully conditional specification (FCS) or multiple imputation using chained equations (MICE) is a flexible alternative to joint models. The procedure specifies an individual regression model for each variable using the other variables in the model as predictors.

Empirical evidence has suggested that MI is unbiased when the data are normally distributed (e.g., Choi, Golder, Gillmore, & Morrison, 2005; Collins, Schafer & Kam, 2001; Graham & Schafer, 1999; Raghunathan, Lepkowski, Van Hoewyk, & Solenberger, 2001; Van

Buuren, Brand, Groothuis-Oudshoorn, & Rubin, 2006), but its performance under nonnormality is unclear. A simulation by Demirtas, Freels and Yucel (2008) showed that MI performed reasonably well when the normality assumption was clearly violated and the sample was relatively large (n > 400). According to Van Buuren (2012), the effect of nonnormality is generally small for measures that rely on the center of the distribution but could be substantial for other types of estimates. A simulation involving several sequential regression MI methods (and extensions which adjust for nonnormal error terms) found that MI performed well for estimating marginal means and regression coefficients when the error distribution was flat or moderately heavy tailed but had poor performance when the distribution was strongly heavy tailed (He and Raghunathan, 2009).

Although MI has been successfully used in large epidemiologic and biomedical datasets (e.g., Centers for Disease Control and Prevention AIDS surveillance system, National Health and Nutrition Examination Survey, National Medical Expenditure Survey), its small sample properties are unclear. Early research by Graham & Schafer (1999) showed that MI performed well when samples were small (n = 50), but recent simulations suggest that MI may have biases in small samples, even when the data are normally distributed (Demirtas et al., 2008; Von Hippel, 2013b).

Given its potential for improving the validity of RCT results (Sterne et al., 2009), there has been a call for investigations into MI's properties and limitations (Enders, 2010; Graham, 2012; Lee & Carlin, 2012; Rässler, Rubin & Zell, 2013; Stuart, Azur, Frangakis, & Leaf, 2009). In addition to an appeal for more systematic research, there have been specific requests for investigations into MI's effect on small samples (Von Hippel, 2004) and power (Davey & Savla, 2010; Young, Weckman & Holland, 2011). In a call to action, the National Academy of Science,

in a special report to the Food and Drug Administration (2010), urged sponsors of clinical trials to make the treatment of missing data a priority, proposed that approved missing data techniques be limited to those that account for the uncertainty attributable to missing data (e.g., MI), and identified several high priority areas for missing data research (e.g., the effect of missing data on power, the robustness of missing data methods, and the development of software that supports coherent missing data analysis).

In response to demands for more principled methods of handling missing data, IBM SPSS Statistics (hereafter referred to as SPSS) added easy to implement MI routines based on the chained equation approach (see Grace & Sawilowsky, 2009, for a comparison of missing data software). By eliminating the need for specialized software and advanced analytical training, the program allows clinicians, unfamiliar with MI, to utilize the technique when analyzing incomplete data.

Statement of the Problem

Methodologists have described missing data as "one of the most important statistical and design problems in research" (methodologist William Shadish, cited in Azar, 2002, p. 70). The National Academy of Science (2010) identified numerous high priority areas for missing data research that are echoed in the literature. This study will address several of those areas by systematically investigating the impact of MI on the rejection rate of the independent samples *t* test (also referred to as the *t* test) under a range of conditions that reflect the interplay of complexities that arise when analyzing differences between treatment arms in RCTs. More specifically, this investigation will utilize a factorial design that will examine eight sample sizes, five treatment effect sizes, three fractions of missing data, three distributions, and two alpha levels to determine if MI impacts Type II error in RCTs.

In addition to identifying areas for missing data research, the National Academy of Science (2010) urged analysts and clinical reviewers to become familiar with current missing data terminology and techniques. To help clinicians meet this objective, this study also seeks to provide an overview of the MI procedure, as implemented in SPSS, with a focus on the practical aspects and challenges of using this method.

Human Participants

This research will utilize a Monte Carlo simulation approach that does not involve human subjects.

Declaration of Interest

The author has no competing interests.

Table 1

Abbreviations & Symbols

Abbreviation or Symbol	Definition
α	The probability of making a Type I error
β	The probability of making a Type II error (1- β denotes statistical power)
CI	Confidence interval
d	Effect size (Cohen's **d**)
df	Degrees of freedom
ES	Effect size
FCS	Fully conditional specification
FMI (γ)	Fraction of missing information
HRR	Highest rejection rate
ITT	Intention to treat
λ	Proportion of variance attributable to missing data
LD	Listwise deletion (also referred to as complete case analysis)
LRR	Lowest rejection rate
MAR	Missing at random
MCAR	Missing completely at random
MICE	Multiple imputation by chained equations
MIS	Missing
MS	Mean substitution
MNAR	Missing not at random
MI	Multiple imputation
μ	Population mean
n	Number of cases (generally in a subsample)
OBS	Observed
PD	Pairwise deletion
PDI	Percentage of data imputed
θ	Parameter of interest
RCT	Randomized controlled trial
RE	Relative efficiency
RI	Regression imputation
σ	Population standard deviation
SPSS	IBM SPSS Statistics
t test	Independent samples *t* test

CHAPTER 2 LITERATURE REVIEW

Because RCTs create equivalent groups (i.e., balance prognostic factors) by randomly assigning participants to different treatment arms, they allow researchers to draw causal conclusions about the efficacy and adverse effects of treatments or interventions. Less than optimum strategies for treating missing values can undermine random assignment and seriously compromise the validity of clinical trials. MI is a promising alternative to traditional methods that has been shown to be unbiased when the data are normally distributed (e.g., Choi, et al., 2005; Collins, et al., 2001; Graham & Schafer, 1999; Raghunathan, et al., 2001; Van Buuren et al., 2006), but its small sample properties and performance under nonnormality are unclear. The purpose of this study is to investigate the impact of MI on the rejection rate of the independent samples *t* test under varying conditions of sample size, effect size, fraction of missing data, distribution shape, and alpha. To provide readers with a solid understanding of the different facets of this research, this chapter is divided into three sections. The first section provides an overview of missing data, including its theoretical underpinnings. More specifically, it examines missing data mechanisms, patterns of missing data, traditional approaches to the handling of missing data, and the pervasiveness of traditional approaches. The second section provides a foundation for understanding the outcome of this study. It explains statistical power, the almost total lack of attention to Type II error in the literature, and the impact of Type II error on the validity of RCTs. The final section reviews the results of previous investigations into the performance of MI; describes the MI procedure, as implemented in SPSS, with a focus on the practical aspects and challenges of using this method; and examines the robustness of the statistical test that will be used in the analysis phase of the MI procedure.

Overview of Missing Data

Missing data can seriously affect the validity of clinical research (McKnight, McKnight, Sidani, & Figueredo, 2007). The magnitude of the impact depends on the mechanisms that led to the missing data, the amount of missing data, and the pattern of missing data (McKnight, et. al, 2007; Tabachnick & Fidell, 2007).

Mechanisms. Rubin (1976) identified three missing data mechanisms that serve as probabilistic explanations for how missing values are related to variables in a dataset. The mechanisms are not characteristics of the dataset but rather assumptions that apply to specific analyses. Although it is often difficult to discern the form of mechanisms (Collins et al., 2001), a sensitivity analysis conducted by Graham, Hofer, Donaldson, Mackinnon, and Schafer (1997) showed that the effects of inaccessible mechanisms are minimal in the implementation of MI. The mechanisms can be classified as missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR).

Although MCAR is a strict assumption that is rarely satisfied in practical applications (Acock, 2005; Enders, 2001; Raghunathan, 2004), it is the principal assumption underlying the traditional approaches to missing data. Under the MCAR condition, missing values are not related to either observed or missing values in a dataset. Because the distribution of missing values cannot be predicted, the observed data can be treated as a simple random sample and the mechanism capturing the reason for the missing values can be ignored for sampling-based and likelihood-based inferences (Little & Rubin, 2002). Using Rubin's (1976) notation, the MCAR mechanism can be represented by

$$p\,(\mathrm{R}|\theta)$$

where *p* is the probability distribution, R is the response indicator (observed or missing), and θ is a parameter that describes the relationship between R and the data.

MCAR is the only missing data mechanism that is testable. Although it does not provide definitive evidence, Little's (1988) multivariate test can assess whether the MCAR condition is tenable (Littell, Roderick, & Schenker, 1995). The test produces a chi-square value that compares the observed variable means for each pattern of missing data with the expected population means. If the test is significant, the data are not MCAR. Although MCAR is not an assumption of MI, the test is still useful for identifying correlates of missingness that should be included in the imputation model. MI has been shown to be unbiased and efficient under the MCAR condition (Enders, 2006).

The MAR mechanism is less restrictive and more tenable than MCAR and is the principal assumption required for most forms of imputation. Under this condition, missing values are related to observed values, but not to missing values. The MAR mechanism can be represented by

$$p\,(R|Y_{obs}, \theta)$$

where *p* is the probability distribution, R is the response indicator (observed or missing), Y_{obs} is the observed data, and θ is a parameter that describes the relationship between R and the data. MAR is the most common mechanism in epidemiologic research (Moons, Donders, Stijnen, & Harrell, 2006) and MI has been shown to be unbiased and efficient under this condition (Buhi, Goodson, & Neilands, 2008; Little & Rubin, 2002). Because MI does not require information about θ, the mechanism capturing the reason for missing data can be ignored.

The MNAR condition is present when missing values are related to the values that are missing and cannot be ignored. The mechanism can be represented by

$$p\,(R|Y_{obs}, Y_{mis}, \theta)$$

where *p* is the probability distribution, R is the response indicator (observed or missing), Y_{obs} is the observed data, Y_{mis} is the missing data, and θ is a parameter that describes the relationship between R and the data. MI has been shown to perform reasonably well under the MNAR condition (Collins et. al, 2001; Sinharay, Stern, & Russell, 2001), even with 25% of the data missing (Buhi et al., 2008). Using a nonignorable imputation model (Van Buuren, 2012) or including variables that account for missingness in the imputation model (Moons et al., 2006; Van Buuren, 2012) can reduce the impact of MNAR. Readers who are interested in MNAR analysis methods are referred to Enders, 2010.

Amount of missing data. There is not a consensus regarding the percentage of missing data that can be tolerated by RCTs. Even a few MNAR values can seriously affect the generalizability of results (Tabachnick & Fidell, 2007). Evidence suggests that the bias introduced by the fraction of missing data may depend on the method used to address the problem. A simulation by Berlin (2009) demonstrated that listwise deletion could incorrectly yield a nonsignificant treatment effect (i.e., a Type II error) with only 5% of the data missing. In contrast, Choi et al. (2005) showed that MI (as employed in the statistical programs EMCOV, NORM, Amos, and Mplus) could provide parameter estimates that come close to those of the population with 50% of the data missing.

Missing data patterns. The MI procedure recognizes two patterns of missing data: monotone and arbitrary. Readers who are interested in other missing data patterns (e.g., latent variable, missing by design) are directed to Enders, 2010; and Little & Rubin, 2002. A monotone pattern (Figure 1) is typically associated with attrition (Enders, 2010; Rässler et al., 2013) and is present when data on an individual measurement are missing on all subsequent measurements.

Let Y_j be the *j*th variable, $j = 1, 2, \ldots, p$ in a dataset. Having a missing value on variable Y_j also means having missing values on all of the following variables $Y_{j+1}, \ldots, Y_p$. The univariate pattern (Figure 2) is a special type of monotone pattern that occurs when missing values are confined to a single variable. This pattern frequently occurs in experimental studies (Enders, 2010). If the data follow a monotone pattern, SPSS imputes the missing values using a noniterative estimation algorithm (see Schafer, 1997, p. 218 – 238 for discussion). The arbitrary pattern (Figure 3) is present when missing values are randomly dispersed throughout the data matrix. If the data follow this pattern, SPSS uses an iterative Markov chain Monte Carlo method to impute the missing values.

Number of Cases	Missing Patterns[a]								Complete if …[b]
	VAR001	VAR002	VAR003	VAR004	VAR005	VAR006	VAR007	VAR008	
35									35
5								X	40
5							X	X	45
5						X	X	X	50
5					X	X	X	X	55
5				X	X	X	X	X	60

a. Variables are sorted on missing patterns.
b. Number of complete cases if variables missing in that pattern (marked with X) are not used.

Figure 1. Monotone pattern of missing data. Simulated SPSS output.

Number of Cases	Missing Patterns[a]								Complete if …[b]
	VAR001	VAR002	VAR003	VAR004	VAR005	VAR006	VAR007	VAR008	
45									45
15								X	60

a. Variables are sorted on missing patterns.
b. Number of complete cases if variables missing in that pattern (marked with X) are not used.

Figure 2. Univariate pattern of missing data. Simulated SPSS output

Number of Cases	Missing Patterns[a]								Complete if …[b]
	VAR001	VAR002	VAR005	VAR007	VAR003	VAR006	VAR004	VAR008	
44									44
3						X			47
4								X	48
1					X			X	50
1					X				45
1				X	X				47
1				X					45
3							X		47
1			X				X		49
1			X						45

a. Variables are sorted on missing patterns.
b. Number of complete cases if variables missing in that pattern (marked with X) are not used

Figure 3. Arbitrary pattern of missing data. Simulated SPSS output.

Traditional Approaches to Missing Data

Missing data threaten the validity of clinical trials, yet receive little attention in the literature (O'Neill & Temple, 2012; Wood et al., 2004). Poor approaches for treating missing values can produce biased estimates, distorted statistical power, and invalid conclusions (Acock, 2005; Enders, 2010; Fichman & Cummings, 2003; Graham, 2012). These consequences are especially serious for phase III trials, which are intended to provide evidence of the efficacy and safety of medical treatments. The use of inadequate missing data methods can also impede the construction of valid prognostic models (Burton & Altman, 2004), undermine random assignment in RCTs, and violate the ITT approach to the analysis of clinical trial data.

Given their potential to compromise inferences from RCTs, a brief overview of traditional techniques is warranted. Although a multitude of approaches have been proposed in the literature, the focus on cross-sectional designs precludes an exhaustive review of these methods. Readers who are interested in procedures that are utilized in repeated measures designs (e.g., last observation carried forward) are directed to Allison, 2001; Little and Rubin, 2002; and

Schafer and Graham, 2002. Although the following techniques are supported in SPSS, the literature is replete with illustrations that show the detrimental effects these approaches have on results (e.g., Acock, 2005; Allison, 2001; Baraldi & Enders, 2010; Enders, 2010; Enders & Bandalos, 2001; Graham & Schafer, 2002; Olinsky, Chen, & Harlow, 2003; Raghunathan, 2004; Schafer & Graham, 2002).

Listwise deletion (LD) is also known as complete case analysis and is the default in SPSS (and most other standard statistical packages). Although Graham and Donaldson (1993) describe special cases where LD did not bias estimates under MAR, most of the literature confirms that LD biases estimates while underestimating variances, covariances, and correlations when the data do not meet the MCAR assumption (e.g., Enders & Bandalos, 2001; Graham & Schafer, 2002; Von Hippel, 2004). Because LD discards every case that has one or more missing values, it can drastically reduce sample size. According to Acock (2005), LD typically results in discarding 20% to 50% of the data (p. 1015). This loss substantially increases the risk of a Type II error and the reduction in power can be devastating, even for large samples with relatively small amounts of missing data. Choi et al. (2005) illustrated how the use of LD completely obliterates power. In their demonstration, the authors showed how removing 20% of the data reduced their sample from 463 to 32 and how removing 50% of the data reduced their sample from 463 to 1. As the nondiscarded cases may not be representative of the population, LD undermines external validity (Allison, 2001; Enders, 2010; Schafer & Graham, 2002) and violates the ITT approach to the analysis of clinical trial data.

Pairwise deletion (PD) is also known as available case analysis and uses the observed data for each analysis without attempting to restore the rectangular form of the data matrix (which in some procedures, like structural equation modeling, may prevent a solution).

According to McKnight et al. (2007), “large discrepancies between the number of available cases for each of the variables in the analysis often produce interpretation problems that are insurmountable” (p. 99). Because the appropriate degrees of freedom for tests of significance are ambiguous, the estimated standard errors and test statistics produced by conventional software are biased and tend to increase Type II errors (Allison, 2001). In models involving only one or two variables (e.g., *t* test, one way ANOVA), LD and PD are identical methods.

Mean substitution (MS) replaces missing values with the mean of observed values. According to Acock (2005), “the mean substitution approach is probably the worst solution to missing values because it attenuates variance and often provides poor imputed values” (p. 1025). Because each imputed value falls directly on a straight line with a slope of zero, MS can dramatically attenuate correlations between variables (Baraldi & Enders, 2010). The appreciable bias observed when MS is used should cause concern because it can lead to a substantial reduction in power (Schlomer, Bauman, & Card, 2010). Schlomer et al. (2010) illustrated how the use of MS would have led to the incorrect conclusion that there was no difference between two treatments when there was actually a large difference between the arms.

Regression imputation (RI) replaces missing values with predicted values obtained from a linear regression equation without incorporating a stochastic component to account for uncertainty. Consequently, RI underestimates variance and lacks the variability that would be present in the hypothetically complete dataset. Because imputed values fall directly on a straight line with a nonzero slope, RI overestimates correlations (Baraldi & Enders, 2010), even when the data are MCAR (Enders, 2010).

Despite the fact that Wilkinson and the Task Force on Statistical Inference (1999) declared LD and PD “among the worst methods available for practical applications” (p. 598),

these techniques continue to dominate the literature – even in high-impact medical journals with stringent statistical review policies. Wood et al. (2004), for example, reviewed 71 RCTs published between July 2001 and December 2001 in four prestigious medical journals (i.e., Journal of the American Medical Association, British Medical Journal, The Lancet and New England Journal of Medicine). The authors found that LD was used in 92% of the cross-sectional and 46% of the repeated measures designs, and noted that the use of this technique has the potential to cause substantial bias in treatment effect estimates. The findings of Peugh & Enders (2004) also corroborate the popularity of LD. In their review of 23 applied research journals published in 1999 and 2003, they found that 96% of the studies used LD or PD.

The use of LD also extends to medical research conducted with strict oversight. Harel, Pellowski, and Kalichman (2012) reviewed 57 RCTs maintained by the HIV/AIDS Prevention Research Synthesis Project at the Centers for Disease Control in June 2010 and found that LD was used in 74% of the studies. Under relaxed assumptions, the authors "expect only 12% of the studies to report unbiased results" (p. 1382).

Klebanoff and Cole (2008) attempted to document the use of MI procedures appearing in the epidemiologic literature (i.e., American Journal of Epidemiology, Annals of Epidemiology, Epidemiology, and International Journal of Epidemiology) from January 2005 to December 2006, but the rarity of MI use precluded analysis. In a similar attempt to document the transition from traditional methods to more principled methods, Mackinnon (2010) recorded the number of MI studies appearing in four medical journals (i.e., Journal of the American Medical Association, New England Journal of Medicine, British Medical Journal, and The Lancet) at two time points (before 2005 and from 2005 to 2008). Although Mackinnon reported that the use of MI in clinical trials has "risen substantially" (2010, p. 586), the increase was based on an

extremely small number of publications (e.g., an increase from 1 RCT appearing in the New England Journal of Medicine before 2005 to 4 appearing from 2005 to 2008).

The Lack of Attention to Power in the Literature

Statistical power (rejecting a false null hypothesis) has received less attention than it deserves (Cohen, 1962, 1988, 1990; Murphy et al., 2009). The almost total lack of attention to Type II error (failure to reject a false null hypothesis) and its consequences have "worrying implications" (Williams & Seed, 1992, p. 321). In clinical research, a high probability of Type II error can lead to the underreporting of serious adverse events (e.g., death, major bleeding, serious infections) and erroneous conclusions of equivalent toxicity (Tsang et al., 2009). The conclusions derived from underpowered studies are often contradictory (Howard, et al., 2000; Maxwell, 2004; Rossi, 1990) and make it difficult to draw coherent clinical inferences from the literature (Button et al., 2013; Maxwell, 2004). Not only do underpowered studies lead to a confusing literature, they also adversely affect future research by creating a reference literature that contains biased effect size estimates (Maxwell, 2004). Failing to detect the effects of treatments or interventions may also contribute to the premature termination of potentially valuable research (Cohen, 1962; Williams & Seed, 1993; Yuen & Pope, 2008) and this is especially true when Type II errors are committed in exploratory studies involving innovative designs or small treatment effects (Chase & Chase, 1976; Freiman et al., 1978; Woods et al., 2006). Rosenthal (1990) provided an example that demonstrates how aspirin would have been deemed ineffective in preventing heart attacks (with an r^2 = .001) if the trial had not been sufficiently powered.

Although the literature is replete with warnings about the potentially disastrous impact of underpowered studies, recommendations by Wilkinson and the Task Force on Statistical

Inference (1999) to improve statistical practice have gone unheeded. A review of the literature showed that RCTs are often inadequately powered to detect iatrogenic effects.

In a trial comparing fixed-dose, weight adjusted unfractionated heparin with low molecular weight heparin for the treatment of venous thromboembolism, Kearon et al. (2006) failed to find a significant difference in the frequency of major bleeding events despite observing twice as many events in the low molecular weight group (12 out of 352) than in the unfractionated group (6 out of 345). A subsequent evaluation of the trial revealed that the power to detect the difference in the proportions was .30 (Tsang et al., 2009).

The failure of Kearon et al. (2006) to detect adverse effects is not uncommon. In an investigation into the cognitive effects of subthalamic nucleus deep brain stimulation in Parkinson's disease, Woods et al. (2006) reviewed 30 studies published between 1997 and 2004 and found that only 7% of the studies demonstrated adequate power ($\geq$ .80) to detect the cognitive decline associated with large (f = .40) effects.

In a similar meta-analysis, Tsang et al. (2009) reviewed six RCTs published between January 2006 and March 2007 to determine if RCTs were sufficiently powered to detect serious adverse events. Their results revealed statistical power levels that ranged from .07 to .37. The authors noted that erroneous conclusions of equivalent efficacy and toxicity were being drawn (p. 610).

A review of the literature has also shown that RCTs are often inadequately powered to detect treatment effects. Yuen & Pope (2008) investigated the power of RCTs in the treatment of non-renal SLE. Their review of 30 negative trials published between 1975 and 2007 revealed a mean statistical power of .25. The authors found that only one of the RCTs demonstrated

adequate power (≥ .80). They concluded that useful therapies could be discarded (p. 1369) and that "the generalizability of SLE trials was modest at best" (p. 1370).

Pike & Leith (2009) also found that RCTs are underpowered. They examined 29 superiority trials appearing in the orthopedic literature from 1994 to 2007. Their results revealed a mean statistical power of .41. The authors concluded that none of the trials were sufficiently powered to detect a small treatment effect, two (6.9%) were sufficiently powered to detect a medium treatment effect, and 13 (44.8%) were sufficiently powered to detect a large treatment effect.

In a comprehensive examination of power in the neuroscience field, Button et al. (2013) reviewed 49 meta-analyses (comprised of 730 primary studies) published in 2011 and found that the median statistical power was .21. The authors concluded that "there is now substantial evidence that a large proportion of the evidence reported in the scientific literature may be unreliable" (p. 374).

In an older investigation of Type II error, Brown, Kelon, Ashton, and Werman (1987) examined 13 negative RCTs appearing in the emergency medicine literature from 1972 to 1984 and found that the statistical power ranged from .03 to .40. They noted that for the endpoints examined, a sample size of up to 450 times larger than that used would have been required to detect a clinically important difference. According to the authors, "this raises serious ethical issues because study subjects were enrolled in a trial that at the outset was doomed to be negative" (p. 187).

To assess changes in statistical practice, Moher et al. (1994) compared 102 negative RCTs published over a 20-year period (1975, 1980, 1985, and 1990) and concluded that the

statistical power of RCTs has not improved over time. Current meta-analyses (e.g., Button et. al., 2013; Yuen & Pope, 2008) collaborate Moher et al.'s (1994) findings.

The prevalence of underpowered RCTs in the literature may result from a misunderstanding of alpha, beta, and power by the clinical community (Cohen, 1992). To clarify, a Type I error (α) is the probability of finding a statistical difference between treatment arms when the treatments are equivalent (rejecting a true null hypothesis). A Type II error (β) is the probability of not finding a statistical difference between treatment arms when the treatments are not equivalent (failure to reject a false null hypothesis). Power is the probability of detecting a difference between treatment arms when the treatments are not equivalent (rejecting a false null hypothesis) and can be defined as $1 - \beta$.

Power is a function of alpha (α), sample size (*n*), and effect size (ES). Increasing any of these parameters will increase the power of a statistical test. Cohen (1965) recommended that power = .80 ($\beta = .20$) when $\alpha = .05$. This proposes a 4:1 β:α ratio. According to Cohen (1992), "a materially smaller value than .80 would incur too great a risk of a Type II error. A materially larger value would result in a demand for *n* that is likely to exceed the investigator's resources" (p. 156). The following illustrations, created with G*Power 3 (Faul, Erdfelder, Lang and Buchner, 2007), show how reducing n_1, n_2 from 60, 60 (Figure 4) to 10, 10 (Figure 5), at alpha .01, increases the probability of committing a Type II error (β) from 4% to 83% when a large (0.80) treatment effect is present.

Cohen (1988) suggested a number of conventions for describing treatment effects as small, medium or large. The recommended effect sizes (**d**) for the independent samples *t* test are 0.2σ for a small effect, 0.5σ for a medium effect, and 0.8σ for a large effect. Sawilowsky (2009)

extended Cohen's work to include very small (0.01σ), very large (1.2σ), and huge (2.0σ) effect sizes.

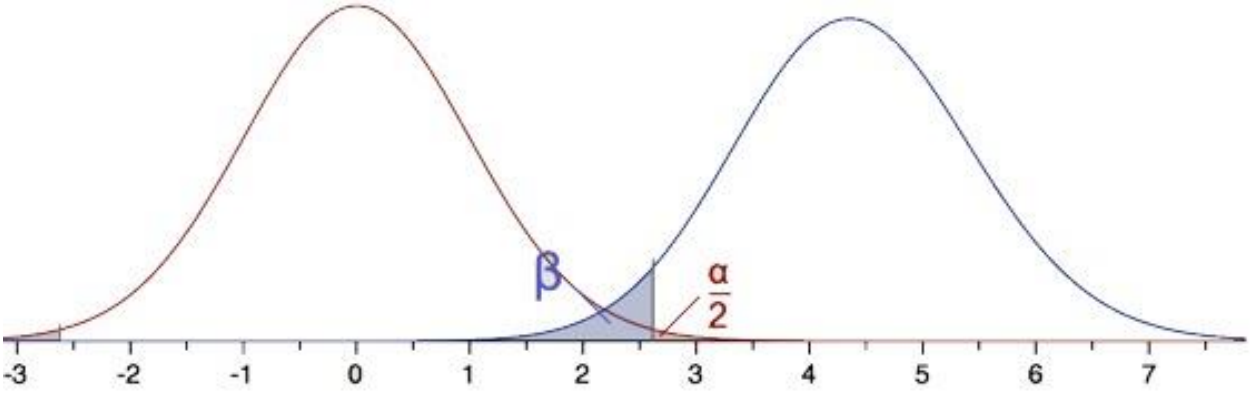


Figure 4. Two-tailed t test, n_1, n_2 = 60, 60; **d** = 0.80; α = .01, power = .959

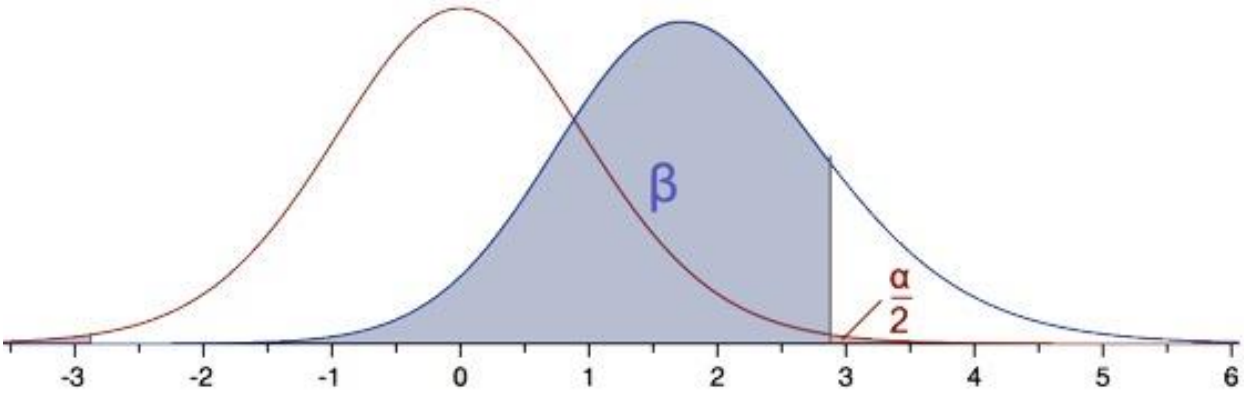


Figure 5. Two-tailed t test, n_1, n_2 = 10, 10; **d** = 0.80; α = .01, power = .172

Multiple Imputation

Multiple imputation (MI) is a promising approach to treating missing data. First proposed by Rubin (1976) and elaborated in 1987, MI replaces missing values with $m > 1$ sets of imputed values, resulting in m complete datasets. Each of the datasets is analyzed and the results are combined to yield one set that reflects both within- and between-imputation uncertainty. By accounting for the variability between imputations, MI allows the uncertainty in the imputation

process to be quantified and integrated into the analysis, which allows it to provide accurate inferential conclusions (Schlomer et al., 2010).

A plethora of simulation studies show MI estimates to be unbiased (e.g., Choi et al., 2005; Collins, et al., 2001; Graham & Schafer, 1999; Raghunathan et al., 2001; Van Buuren et al., 2006). A simulation by Choi et al. (2005) showed that MI provided parameter estimates that came close to those of the population with 50% of the data missing. These results were corroborated by Marshall, Altman, Royston, and Holder (2010) who also reported MI to be useful when up to 50% of the values were missing. Although these results are encouraging, there may be an upper limit to the amount of missing data that can be tolerated by MI. Barzi and Woodward (2004) revealed inflated variability and convergence problems with some MI techniques when more than 60% of the observations had missing data (with varying results depending upon the imputation technique).

Nonnormality. MI procedures are somewhat robust against violations of normality (as far as bias is concerned). According to Van Buuren (2012), the effect of nonnormality is generally small for measures that rely on the center of the distribution but could be substantial for other types of estimates. Demirtas et al. (2008) found that with a fairly large sample ($n > 400$), MI performed reasonably well when the normality assumption was clearly violated (i.e., flatness of the density, heavy tails, non-zero peakedness, skewness, and multimodality), even when there was a high percentage (75%) of missing data. Graham and Schafer (1999) conducted a simulation in which highly nonnormal variables were imputed under normality assumptions with no transformations or rounding and reported excellent performance for linear regression. Von Hippel (2013a) imputed skewed variables from a normal model and found that they produced acceptable estimates for means, variances, and regressions. A simulation of several

sequential regression MI methods (and extensions which adjust for nonnormal error terms) found that when the error distribution was flat or moderately heavy tailed, MI was able to estimate marginal means and regression coefficients, but when the distribution was strongly heavy tailed, the regression coefficients were biased (He & Raghunathan, 2009).

Despite MI's potential for poor performance under conditions of extreme skew, strong concerns over the use of transformations (in the context of MI) have been raised (e.g., Demirtas et al., 2008, p. 82 – 83). After testing several methods for adapting a normal imputation model to accommodate skew (e.g., transform, truncate or round), Von Hippel (2013a) found that none of the modifications reliably reduced bias (and some modifications made the bias much worse). Enders (2010) pointed out that because MICE relies on the associations among variables and transformations can alter the covariate structure of the data, transforming (then back transforming) the data could affect the accuracy of the imputations and resulting estimates.

MNAR. MI has been shown to perform reasonably well under the MNAR condition (Collins et al., 2001; Sinharay et al., 2001; Carpenter, Kenward, & White, 2007) with a moderate amount (25%) of missing data (Buhi et al., 2008; Collins, et al., 2001). Under more extreme conditions (i.e., the missing data rate exceeds 25% or in the case of a linear regression $r > .40$), the form of the mechanism (whether the probability to be missing was linear or was more likely to occur in the extremes) determined which parameters were affected (Collins et al., 2001). The impact of MNAR can be reduced by including auxiliary variables that account for missingness in the imputation model (Moons et al, 2006; Van Buuren, 2012) or by using a nonignorable imputation model (Van Buuren, 2012).

Small Samples. Although MI has been successfully used in large epidemiologic and biomedical datasets (e.g., Centers for Disease Control and Prevention AIDS surveillance system,

National Health and Nutrition Examination Survey, National Medical Expenditure Survey), its small sample properties are unclear. Graham and Schafer (1999) showed that MI performed well when samples were small (n = 50) and a large proportion (50%) of the data were missing. In fact, some of the analyses based on MI data were as good as the same analyses performed on complete data (Graham & Schafer, 1999). Barnes, Lindborg and Seaman (2006) showed that common regression based MI methods provided close to ideal CI coverage for 20% dropout in small (n = 20, 30, 50) clinical trials, but some methods fell short when the percentage of missing data increased to 30 or 40%. Von Hippel (2013b) demonstrated that bias occurred when the samples were small (n = 25), there was a large amount of missing data (50%), and the missing values followed an exceptionally challenging pattern. Demirtas et al. (2008) also showed that MI produced biased estimates when samples were small (n = 40) and a large amount (75%) of data were missing. Simulations by Kim (2004) showed that decreasing the sample size from 200 to 20, increased the variance of point estimators by a factor of 10 or more depending upon the proportion of missing data.

Multiple Imputation in Practice

Initial multiple imputation procedures assumed a large joint model (e.g., a joint normal distribution). As almost all datasets have mixtures of incomplete categorical and continuous variables, this assumption rarely holds in practice. MICE is a flexible alternative to joint models that does not assume that the data have an underlying normal distribution (Johnson & Young, 2011; Schafer, 1999; Van Buuren, 2012). It also does not assume that nonresponse is ignorable (Schafer, 1999). In principal, imputations can be created under any missing data mechanism and the inferences will be valid under that mechanism (Schafer, 1999). The MI procedure has three phases: imputation, analysis, and pooling.

The imputation phase. During this phase, the MI procedure creates several complete versions of the dataset by iteratively replacing missing values with imputed values. The MICE approach accomplishes this by specifying an individual regression model for each variable using the other variables in the model as predictors $p\ (Y_{jmis} \mid Y_{jobs}, Y_{-j}, R)$. In contrast to joint modeling, MICE specifies the multivariate distribution $p\ (Y, X, R \mid \theta)$ through a set of conditional densities $p\ (Y_j \mid X, Y_{-j}, R\phi_j)$. The conditional density is used to impute Y_j given X, Y_{-j} and R which allows each variable to be modeled according to its distribution. SPSS uses logistic regression to impute incomplete binary and categorical variables, and linear regression or predictive mean matching (a variant of linear regression that matches imputed values computed by the regression model to the closest observed value) to impute continuous variables.

There are several ways to implement imputation under conditionally specified models. The MICE algorithm starts with simple random draws from the marginal distribution and sequentially imputes each variable in the order specified in the variable list (e.g., by missing value rates) until the iteration is complete and all of the variables have been imputed. The process repeats using the Gibbs sampling procedure (a Bayesian simulation technique that samples from the conditional distributions in order to obtain samples from the joint distribution) for a specified number of iterations (in SPSS, the default is 10). A number of simulations have shown that unbiased estimates and appropriate coverage is obtained after 5 – 10 iterations (Raghunathan et al., 2001; Van Buuren, 2012; Van Buuren & Groothuis-Oudshoorn, 1999; White, Royston, & Wood, 2011). When the specified number of iterations has been reached, the distribution of parameters governing the imputations should have converged. At convergence, the complete dataset is retained and the entire process is repeated *m* times resulting in *m* complete datasets being stacked with the original incomplete dataset in the SPSS data file.

When choosing variables to include in the imputation model, a simulation by Collins et al. (2001) showed that an inclusive strategy is superior to a restrictive strategy "because there appear to be few risks associated with it and potentially substantial gains" (p. 350). Although there has been some disagreement on the subject (see Hardt, Herke & Leonhart, 2012), the prevailing view is that the imputation model should include: (a) target variables (variables that will be used in the analysis phase); (b) auxiliary variables (variables intended to predict missingness or improve the model, including variables that preserve correlations and interactions; Collins et al., 2001; Piggot, 2001; Rubin, 1996; Schafer & Olsen, 1998; Sinharay et al., 2001); (c) sample variables (variables that describe aspects of clustered, stratified or longitudinal data); and (d) outcome variables (Collins et al., 2001; Enders, 2010; Little, 1992; Moons et al., 2006).

Although it seems counterintuitive, including outcome variables in the imputation model is necessary to reduce bias when imputing predictor variables and does not overestimate the regression coefficients between outcomes and predictors (Moons et al., 2006). Failing to include variables that mediate between outcomes and predictors can also result in bias and loss of power (Collins et al., 2001; Enders, 2010). When interactions are not modeled, the effects of the correlations (and interactions) between the variables will be biased towards zero (Graham, 2009; Sterne et al., 2009). Variables with missing information should also be included in the imputation model. Although this also seems counterintuitive, simulations have shown that auxiliary variables with missing values are nearly as effective in reducing bias as those with no missing values (Enders, 2010). It should be noted, however, that including variables with large fractions of missing information can slow or prevent convergence (Van Buuren & Groothuis-Oudshoorn, 1999).

Although including a diverse set of variables in the imputation model may reduce bias, there seems to be an upper limit to the number of variables that can be modeled. Chained equations break down at a 1:1 ratio of variables to cases, even with small fractions of missing data (Hardt et al., 2012) and can lead to instability (He & Raghunathan, 2009) or cause the program to fail (White et al., 2011). Hardt et al. (2012) suggested that the ratio of variables to cases (with complete data) should not go below 1:3. Van Buuren and Groothuis-Oudshoorn (1999) recommended selecting a suitable subset of data that contains no more than 15 – 25 variables. Enders (2010) suggested that a reasonable goal may be to maximize the squared multiple correlation between auxiliary variables and target variables using as few auxiliary variables as possible. As variables with low correlations (≤ .40) have a negligible affect on power (Enders, 2010), a good strategy may be to exclude auxiliary variables with low correlations and high fractions of missing information.

Rubin (1987) provided a diagnostic measure that estimates the influence of missing data on parameter estimates. Higher fractions of missing information (FMI or γ) represent higher uncertainty about estimates (and their resulting conclusions). The estimated FMI can be defined as

$$\gamma = \frac{r + 2/(df + 3)}{r + 1}$$

where r (the relative increase in variance due to the missing data) is

$$r = \frac{1 + (m^{-1})B}{\bar{U}}$$

and *df* (based on an approximate *t* distribution) is

$$df = (m - 1)[1 + r^{-1}]^{2}$$

Table 2 illustrates the proportional increase in the standard error for different fractions of missing data (γ) given the number of imputations (*m*). The slight decrease in standard error attributable to an increase in *m* suggests that there is little benefit to using $m > 5$.

Table 2

Proportional Increase in Standard Error for m and γ

	γ						
m	0.10	0.20	0.30	0.40	0.50	0.60	0.70
3	1.02	1.03	1.05	1.06	1.08	1.10	1.11
5	1.01	1.02	1.03	1.04	1.05	1.06	1.07
10	1.00	1.01	1.01	1.02	1.02	1.03	1.03
20	1.00	1.00	1.01	1.01	1.01	1.01	1.02

Note: Table was computed using Excel.

Determining the number of imputations to generate can be based on the relative efficiency (RE) desired. Rubin (1987) showed that the efficiency of an estimate based on *m*, relative to one based on an infinite number of imputations, is

$$RE = \left(1 + \frac{\gamma}{m}\right)^{-1}$$

Table 3

Efficiency as a Function of m and γ

	γ								
m	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90
3	0.97	0.94	0.91	0.88	0.86	0.83	0.81	0.79	0.77
5	0.98	0.96	0.94	0.93	0.91	0.89	0.88	0.86	0.85
10	0.99	0.98	0.97	0.96	0.95	0.94	0.93	0.93	0.92
20	1.00	0.99	0.99	0.98	0.98	0.97	0.97	0.96	0.96

Note: Table was computed using Excel.

Table 3 shows the relative efficiency for γ given *m*. Although the table demonstrates that an increase in *m* can be used to compensate for a large γ, it also shows that modest values of *m* result in parameter estimates that are nearly fully efficient. Because MI enlarges the between-imputation variance *B* by a factor *1/m* before calculating the total variance in $T = U + (1 + m^{-1})$ *B*, the classic advice has been that 3 – 5 *m* is adequate (Rubin, 1987; Schafer, 1997; Schafer & Olsen, 1998). Recently, however, there has been some disagreement in the literature.

Von Hippel (2005) evaluated the impact of different fractions of missing data and found that 10 imputations produced a standard error that was 2% larger than an infinite number of imputations when a large percentage (40%) of values were missing. Graham, Olchowski and Gilreath (2007) investigated the effect of *m* on the statistical power of a test to detect a small (< 0.1) effect and suggested that $m > 5$ may be needed for small effect sizes, large fractions of missing data, and small sample sizes. Bodner (2008) systematically explored the variability of the width of the 95% CI, the *p* value and λ (the proportion of variance attributable to the missing data) under various *m* and recommended that *m* be based on the percentage of cases that are incomplete (a conservative estimate of λ). White, Royston and Wood (2011) concurred with Bodner's (2008) suggestion that *m* be at least equal to the percentage of incomplete cases. Van Buuren and Groothuis-Oudshoorn (1999) suggested setting *m* to the average percentage of missing data but added that "the substantive conclusions are unlikely to change as a result of raising *m* beyond $m = 5$" (p. 51). Johnson and Young (2011) corroborate Van Buuren's statement. They demonstrated that $m = 5$ resulted in the same substantive conclusions as $m = 25$, even when auxiliary variables were removed from the model.

Although the benefits of increasing *m* beyond 5 are still being debated, imputing a large number of datasets may not be practical. Imputing a single dataset with a large model or a large

fraction of missing information requires considerable computation time and can prevent convergence (Graham, 2012).

Another consideration when imputing data is whether or not to define constraints. In SPSS, constraints can be used to (a) restrict the range of imputed values, (b) exclude variables from imputation (e.g., variables with large fractions of missing values and variables that have values that are missing by design), and (c) specify rounding rules (see Horton, Lipsitz, & Parzen, 2003, for caveats). In situations where a subset of cases is believed to be inherently different from the rest of the sample, values may be imputed separately using different (split) models (Rubin, 1987). Although the role of the variable can be confined, it is not necessary to specify whether variables are independent (predictor) or dependent (outcome). Unlike methods that exclude cases with missing predictor variables, MI uses observed values to predict missing values without regard to each variable's role in the analysis phase (Enders, 2010).

After a specified number of iterations is reached, the distribution of parameters governing the imputations (e.g., the coefficients in the regression models) should have converged (such that the order in which the variables were imputed no longer matters). Plotting the variable means and standard deviations at each iteration and imputation can help assess model convergence. When convergence is reached, the variance between the sequences will not be larger than the variance within the sequences and the streams will be intermingled without showing any definite trends (Van Buuren, 2012). Nonconvergence can occur when (a) a large number of variables are modeled, (b) a large number of missing values are imputed, (c) analysis variables are left out of the imputation model, or (d) the matrix is not positive definite (Graham, 2012).

The Analysis Phase. In this phase, each imputed dataset is analyzed using standard statistical procedures. As the simulated design for this study is a parallel group RCT, the data

will be analyzed using the *t* test. William Sealy Gosset (Student, 1908) introduced the *t* distribution to allow the probabilities of small samples to be computed when the population standard deviation is unknown. Since its introduction, the *t* test has played an integral role in evaluating the efficacy of medical treatments. In his seminal paper, Gosset introduced the *t* test by comparing the number of hours of sleep patients gained when treated with dextro- and laevo-forms (optical isomers) of the drug hyoscyamine hydrobromide and concluded that the laevo-isomer was more effective. The *t* test has since become the most used procedure for comparing group means in clinical research (Bridge & Sawilowsky, 1999). The *t* test can be defined as

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\left[\frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}\right]\left[\frac{n_1 + n_2}{n_1 n_2}\right]}}$$

with

$$df = n_1 + n_2 - 2$$

Statistical tests, including the *t* test, are considered robust when they maintain a Type I error rate close to the nominal level of significance while maintaining statistical power (Lix, Keselman & Keselman, 1996). The *t* test has two major assumptions: normality and homogeneity of variance.

Although Wilcox (1998) asserted that a violation of the normality assumption could result in a substantial loss of power (see his example where a small departure from normality reduced the power of the *t* test from .96 to .28), Zimmerman (1987) found the *t* test to be scarcely affected by nonnormality. Later work by Sawilowsky and Blair (1992) demonstrated that the *t* test produced power rates very similar to the levels expected from normal curve theory regardless of population shape, sample size, or effect size. Empirical evidence also demonstrates that the *t* test's Type I error rate is maintained at the nominal level when sample sizes are

approximately equal, sample sizes are fairly large (n > 25 according to Boneau, 1960), and tests are two-tailed rather than one-tailed (Boneau, 1960; Huber, 1972; Sawilowsky and Blair, 1992).

In practice, the normality assumption is far from tenable. Datasets in the medical (Yuan & Bentler, 1999) and psychological (Bradley, 1977; Micceri, 1989) literature are likely to be skewed and heavy tailed. In a comprehensive study of the distributional characteristics of large datasets (almost 70% of the datasets involved 1,000 or more cases), Micceri (1989) found that none of the 440 datasets investigated passed the Kolmogrov-Smirnov test of normality. Nearly half (49.1%) of the datasets had at least one heavy tail and almost three quarters (71.6%) were classified as being moderately to extremely asymmetric (some to the point of being exponentially distributed).

Work by Zimmerman (1987) demonstrated that the *t* test is relatively robust to violation of the homogeneity assumption when samples sizes are equal. If sample sizes differ, then inequality of variances can have a pronounced effect on significance levels and on the probability of Type I error (Zimmerman, 1987). In this case, the *t* test either becomes conservative or liberal depending upon the relationship between sample size and population variance.

When the *t* test is performed, SPSS automatically provides the results for Levene's (1960) test for equality of variances. If the test is significant, the variances are not equal. In RCTs, the homoscedasticity assumption is often violated because the treatment group is more likely to experience greater variability in response to the intervention than the control group.

The pooling phase. During this phase, the results of the individual analyses are combined using mathematical rules developed by Rubin (1987). Most of the statistical procedures, available in SPSS, can produce pooled parameter estimates and standard errors for

multiply imputed datasets. SPSS pools output at two levels: naïve and univariate. At the naïve level only the pooled parameter is available. At the univariate level, the pooled parameter and its standard error, test statistic, effective degrees of freedom, *p*-value, CI, and pooling diagnostics (FMI, RE, relative increase in variance) are available. The *t* test procedure, in SPSS, supports mean pooling at the univariate level and *n* pooling at naïve level. Statistics that are not estimators (e.g., likelihood ratio, chi-square, *p*-value) cannot be combined and SPSS will not allow those analyses to pool. According to notation in IBM SPSS 20 Algorithms (2011), the final estimate of Q is simply the average of the individual point estimates and can be defined as

$$\bar{Q} = \frac{1}{m}\sum_{i=1}^{m} \hat{Q}^i$$

where $\hat{Q}$ is the parameter estimate from the *i*th dataset and *m* is the number of imputations. The estimated total variance is

$$T = \bar{U} + \left(1 + \frac{1}{m}\right) B$$

where B (the between imputation variance) is

$$B = \frac{1}{m-1}\sum_{i=1}^{m} \left(\hat{Q}^i - \bar{Q}\right)\left(\hat{Q}^i - \bar{Q}\right)'$$

and $\bar{U}$ (the within imputation variance) is

$$\bar{U} = \frac{1}{m}\sum_{i=1}^{m} U^i$$

with

$$df = (m-1)[1 + r^{-1}]^{\,2}$$

Unlike other statistical contexts, the *df* in MI is not affected by *n*. Barnard and Rubin (1999) noted that the above *df* equation can produce values that are larger than the *df* in the

complete data and developed an adapted version for small samples. There is a void in the MI literature (as well as the SPSS documentation) that discusses this solution and it is not clear whether SPSS makes this adjustment.

It should be noted that missing elements in

$$\left\{\hat{Q}^i, U^i\right\}_{i=1}^{m}$$

are excluded from the calculations.

The *p*-value for testing H_0: Q = Q is

$$p = \Pr(F_{k,v} \geq F)$$

where

$$F = \frac{1}{k}(\bar{Q} - Q_0)^T \tilde{T} - (\bar{Q} - Q_0)$$

$$k = rank(\tilde{T})$$

$$v = \tilde{v}$$

CHAPTER 3 METHODS

The National Academy of Science (2010) identified numerous high priority areas for missing data research that are echoed in the literature. Several of those areas will be addressed by systematically investigating the impact of MI on the rejection rate of the independent samples *t* test under a range of conditions that reflect the interplay of complexities that arise when analyzing differences between treatment arms in RCTs. More specifically, a factorial design that includes eight sample sizes, five effect sizes, three fractions of missing data, three distributions, and two alpha levels will be used to determine if MI impacts Type II error in parallel group RCTs. Fully crossed, this design represents 720 distinct conditions.

Although enormous amounts of computational resources are required to conduct MI simulations (see Hardt et al., 2012, for an example of a simulation of 100 cases that required more than 200 hours of computing time on a 6 physical core PC optimized for simulations), each of the 720 conditions will be replicated 1,000 times.

Data Generation

Two samples (X_1, Y) of sizes (n_1, n_2) = (10, 10), (20, 20), (30, 30), (40, 40), (60, 60), (10, 30), (20, 60), (30, 90) will be drawn from the Normal (0, 1), Chi-square (df = 1) and t (df = 3) distributions. The samples will simulate a parallel group RCT with X_1 serving as the control group and Y serving as the treatment group. As pilot trials can have samples as small as n = 20 (Barnes, Lindborg & Seaman, 2006), the eight sample sizes proposed are intended to reflect the small balanced and unbalanced designs likely to occur in clinical practice. Several of the sample sizes also mirror those used in past simulation studies that investigated the robustness of the *t* test (e.g., Sawilowsky & Blair, 1992; Sawilowsky & Hillman, 1992). A third variable (X_2) with a .50 correlation with Y will serve as an auxiliary variable in the imputation model.

The control sample (X_1) will be generated using the distribution random variable function in SPSS (version 22). The treatment sample (Y) and auxiliary variable (X_2) will be generated via algorithms presented by Headrick and Sawilowsky (2000) to solve the Fleishman (1978) equation for a .50 correlation. The Fleishman method was chosen because it was shown to generate average values of intercorrelations closer to population parameters than competing procedures for skewed distributions and small sample sizes (Headrick & Sawilowsky, 2000).

Table 4

Solutions to the Fleishman Equation

Distribution	μ	σ	γ_1	γ_2	*a*	*b*	*d*
Normal Distribution	0	1	0	0	0	1	0
Chi-square (*df*=1)	1	√2	√8	12	-.5207	.6146	.02007
t (*df*=3)	0	√3	0	17	0	.3938	.17130

Note. γ_1 = skew, γ_2 = kurtosis

First, the constants for each distribution (presented in Table 4) will be used to solve for r_{xy} where

$$r_{xy} = r^2(b^2 + 6bd + 9d^2 + 2a^2r^2 + 6d^2r^4)$$

Then, the SPSS random variable function will be used to generate three random normal variates (Z_1, Z_2, Z_3) which will be used to solve for X_a and Y_a where

$$X_a = r(Z_1) + (Z_2)\sqrt{1 - r^2}$$

and

$$Y_a = r(Z_1) + (Z_3)\sqrt{1 - r^2}$$

After which, X_a and Y_a will be used to solve for X_b and Y_b where

$$X_b = a + bX_1 + (-a)X_1^2 + dX_1^3$$

and

$$Y_b = a + bY_1 + (-a)Y_1^2 + dY_1^3$$

Finally, X_b and Y_b will be transformed back into their distribution metric using

$$X_o = \sigma X + \mu$$

and

$$Y_o = \sigma X + \mu$$

A treatment effect will be will simulated by applying the algorithm $Y_T = Y_O + k\sigma$ where k is equal to a constant treatment effect and σ reflects the standard deviation of the distribution under investigation (Table 4). Proposed ES include small (0.2σ), medium (0.5σ), and large (0.8σ) as suggested by Cohen (1988) as well as very large (1.2σ) and huge (2.0σ) as proposed by Sawilowsky (2009). Despite warnings from Cohen (1988) about these values becoming de facto standards for research, they are proposed to facilitate comparisons across studies and allow for meta-analysis.

Table 5

Number of Values Treated as Missing

	Sample Size (Y_T)					
Fraction of Missing Data	10	20	30	40	60	90
0.10	1	2	3	4	6	9
0.30	3	6	9	12	18	27
0.50	5	10	15	20	30	45

Table 5 presents the number of values (for variable Y_T) that will be defined as missing. To simulate a monotone missing data pattern, 10%, 30%, or 50% of the Y_T values will be deleted using random uniform numbers. If the rank of the generated number is equal to or less than the percentage specified, the data point will be deleted.

Parameters

Missing values will be imputed using SPSS version 22. Parameters will be set to 10 iterations and $m = 5$ which represent the defaults in SPSS. According to Van Buuren (2012), the software defaults are often reasonable. The defaults were also chosen because they are most likely to be used by practitioners unfamiliar with MI. The analyses will be conducted using a two-tailed independent samples *t* test. The combining step will be performed by SPSS using the formulas presented in the literature review. The Python 2.7 programming language will be used to repeat the entire process 1,000 times for each combination of factor levels and report the rejection rates for the .01 and .05 alpha levels

CHAPTER 4 RESULTS

Tables 6 – 10 present the number of rejections out of 1,000 for the Normal distribution under varying sample sizes, effect sizes, fractions of data imputed, and alpha. The Normal distribution has two parameters: a location parameter (μ) and a scale parameter (σ). The distribution was simulated using a location parameter of 0 and a scale parameter of 1.

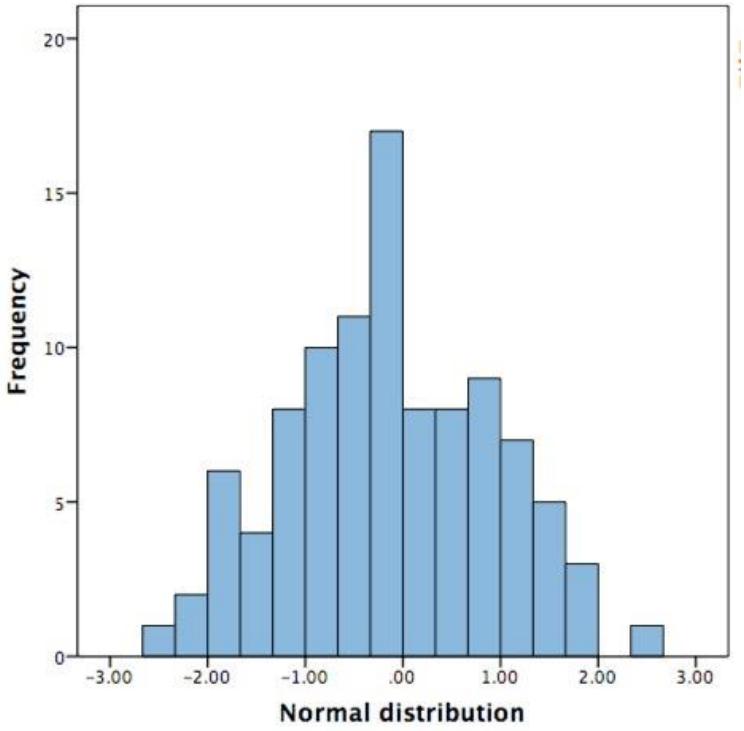


Figure 6. Normal distribution (0, 1). Created using SPSS syntax.

The probability density function for the Normal distribution (Figure 6) is

$$\frac{1}{\sigma\sqrt{2\pi}} e - \frac{(X-\mu)^2}{2\sigma^2}$$

Table 6

Small (0.2σ) Treatment Effect Rejection Rates for the Normal Distribution

	α = .01			α = .05		
	Fraction of data imputed			Fraction of data imputed		
n_1, n_2	0.100	0.300	0.500	0.100	0.300	0.500
10, 10	0.031	0.034	0.023	0.068	0.082	0.064
20, 20	0.028	0.023	0.020	0.103	0.104	0.095
30, 30	0.036	0.032	0.023	0.138	0.107	0.101
40, 40	0.042	0.043	0.034	0.165	0.139	0.116
60, 60	0.067	0.058	0.056	0.195	0.164	0.120
10, 30	0.036	0.019	0.025	0.067	0.084	0.076
20, 60	0.027	0.034	0.039	0.148	0.126	0.100
30, 90	0.048	0.045	0.050	0.137	0.152	0.150

Table 7

Medium (0.5σ) Treatment Effect Rejection Rates for the Normal Distribution

	α = .01			α = .05		
	Fraction of data imputed			Fraction of data imputed		
n_1, n_2	0.100	0.300	0.500	0.100	0.300	0.500
10, 10	0.085	0.067	0.065	0.193	0.179	0.172
20, 20	0.174	0.151	0.107	0.375	0.316	0.251
30, 30	0.231	0.221	0.139	0.429	0.426	0.336
40, 40	0.364	0.264	0.222	0.558	0.535	0.449
60, 60	0.528	0.441	0.350	0.759	0.725	0.599
10, 30	0.114	0.118	0.093	0.266	0.267	0.217
20, 60	0.266	0.235	0.172	0.481	0.449	0.386
30, 90	0.406	0.346	0.309	0.654	0.588	0.576

Table 8

Large (0.8σ) Treatment Effect Rejection Rates for the Normal Distribution

	α = .01			α = .05		
	Fraction of data imputed			Fraction of data imputed		
n_1, n_2	0.100	0.300	0.500	0.100	0.300	0.500
10, 10	0.235	0.183	0.120	0.419	0.355	0.290
20, 20	0.483	0.400	0.284	0.700	0.604	0.514
30, 30	0.684	0.570	0.446	0.839	0.800	0.668
40, 40	0.804	0.740	0.585	0.939	0.882	0.823
60, 60	0.950	0.894	0.788	0.991	0.966	0.938
10, 30	0.348	0.303	0.281	0.582	0.562	0.474
20, 60	0.672	0.637	0.551	0.863	0.842	0.771
30, 90	0.870	0.842	0.757	0.966	0.956	0.918

Table 9

Very Large (1.2σ) Treatment Effect Rejection Rates for the Normal Distribution

	α = .01			α = .05		
	Fraction of data imputed			Fraction of data imputed		
n_1, n_2	0.100	0.300	0.500	0.100	0.300	0.500
10, 10	0.516	0.432	0.320	0.720	0.675	0.555
20, 20	0.878	0.811	0.623	0.954	0.925	0.834
30, 30	0.968	0.923	0.841	0.996	0.984	0.958
40, 40	0.997	0.985	0.935	0.999	0.998	0.991
60, 60	1.000	0.999	0.993	1.000	1.000	1.000
10, 30	0.723	0.692	0.603	0.902	0.862	0.794
20, 60	0.963	0.952	0.925	0.994	0.991	0.968
30, 90	0.999	0.997	0.991	1.000	1.000	0.997

Table 10

Huge (2.0σ) Treatment Effect Rejection Rates for the Normal Distribution

	α = .01			α = .05		
	Fraction of data imputed			Fraction of data imputed		
n_1, n_2	0.100	0.300	0.500	0.100	0.300	0.500
10, 10	0.950	0.887	0.764	0.991	0.967	0.904
20, 20	0.999	0.996	0.979	1.000	1.000	0.999
30, 30	1.000	1.000	0.996	1.000	1.000	1.000
40, 40	1.000	1.000	1.000	1.000	1.000	1.000
60, 60	1.000	1.000	1.000	1.000	1.000	1.000
10, 30	0.996	0.997	0.970	1.000	1.000	0.998
20, 60	1.000	1.000	1.000	1.000	1.000	1.000
30, 90	1.000	1.000	1.000	1.000	1.000	1.000

Tables 11 – 15 present the number of rejections out of 1,000 for the Chi-square (χ^2) distribution under varying sample sizes, effect sizes, fractions of data imputed, and alpha. The Chi-square distribution has a single parameter: *df*. The distribution was simulated using *df* = 1

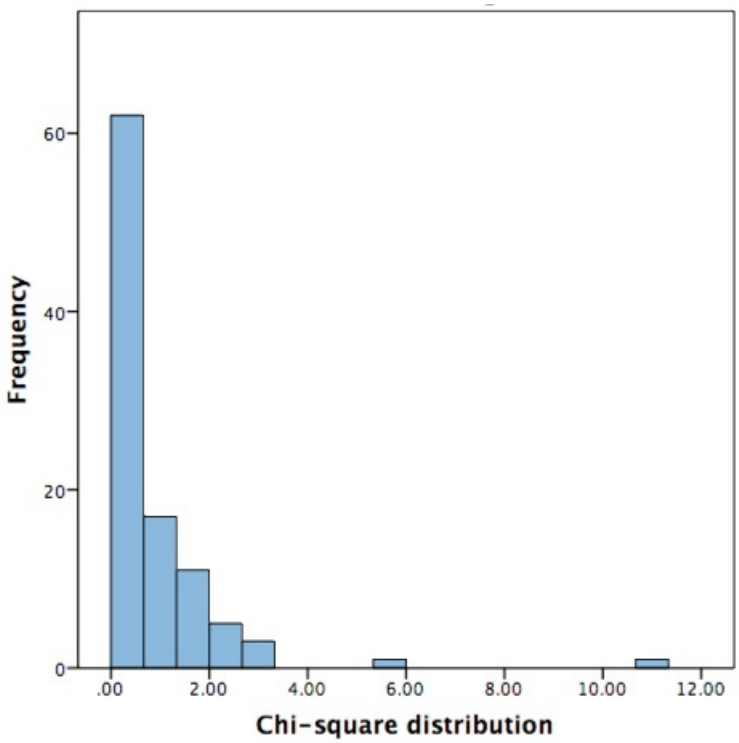


Figure 7. χ^2 distribution (*df* = 1). Created using SPSS syntax.

The probability density function for the χ^2 distribution (Figure 7) is

$$\frac{1}{2\frac{k}{2}\Gamma\left(\frac{k}{2}\right)}x^{\frac{k}{2}-1}e^{-\frac{x}{2}}$$

Table 11

Small (0.2σ) Treatment Effect Rejection Rates for the χ^2 Distribution (df=1)

	α = .01			α = .05		
	Fraction of data imputed			Fraction of data imputed		
n_1, n_2	0.100	0.300	0.500	0.100	0.300	0.500
10, 10	0.034	0.024	0.017	0.093	0.085	0.076
20, 20	0.039	0.033	0.015	0.105	0.117	0.106
30, 30	0.040	0.046	0.046	0.152	0.118	0.122
40, 40	0.053	0.053	0.030	0.160	0.139	0.124
60, 60	0.079	0.067	0.063	0.203	0.167	0.165
10, 30	0.022	0.025	0.017	0.104	0.089	0.084
20, 60	0.033	0.037	0.027	0.134	0.158	0.114
30, 90	0.053	0.054	0.045	0.192	0.162	0.133

Table 12

Medium (0.5σ) Treatment Effect Rejection Rates for the χ^2 Distribution (df=1)

	α = .01			α = .05		
	Fraction of data imputed			Fraction of data imputed		
n_1, n_2	0.100	0.300	0.500	0.100	0.300	0.500
10, 10	0.144	0.115	0.094	0.278	0.250	0.222
20, 20	0.214	0.197	0.168	0.413	0.387	0.316
30, 30	0.317	0.251	0.195	0.515	0.454	0.394
40, 40	0.386	0.361	0.257	0.631	0.569	0.463
60, 60	0.547	0.493	0.372	0.768	0.724	0.609
10, 30	0.147	0.173	0.158	0.354	0.352	0.289
20, 60	0.295	0.281	0.255	0.537	0.508	0.468
30, 90	0.455	0.401	0.366	0.681	0.611	0.580

Table 13

Large (0.8σ) Treatment Effect Rejection Rates for the χ^2 Distribution (df=1)

	α = .01			α = .05		
	Fraction of data imputed			Fraction of data imputed		
n_1, n_2	0.100	0.300	0.500	0.100	0.300	0.500
10, 10	0.325	0.306	0.212	0.520	0.484	0.403
20, 20	0.543	0.470	0.365	0.712	0.664	0.580
30, 30	0.709	0.624	0.469	0.843	0.804	0.705
40, 40	0.819	0.726	0.643	0.914	0.894	0.817
60, 60	0.937	0.898	0.796	0.977	0.957	0.920
10, 30	0.448	0.411	0.375	0.651	0.620	0.564
20, 60	0.683	0.678	0.609	0.864	0.834	0.796
30, 90	0.856	0.853	0.769	0.947	0.925	0.916

Table 14

Very Large (1.2σ) Treatment Effect Rejection Rates for the χ^2 Distribution (df=1)

	α = .01			α = .05		
	Fraction of data imputed			Fraction of data imputed		
n_1, n_2	0.100	0.300	0.500	0.100	0.300	0.500
10, 10	0.636	0.552	0.463	0.754	0.708	0.623
20, 20	0.860	0.797	0.665	0.929	0.894	0.841
30, 30	0.948	0.897	0.821	0.984	0.971	0.950
40, 40	0.991	0.964	0.914	0.994	0.990	0.984
60, 60	0.999	0.994	0.980	1.000	0.999	0.999
10, 30	0.759	0.723	0.663	0.900	0.858	0.830
20, 60	0.925	0.936	0.890	0.987	0.982	0.969
30, 90	0.987	0.988	0.978	0.997	0.996	0.995

Table 15

Huge (2.0σ) Treatment Effect Rejection Rates for the χ^2 Distribution (df=1)

	α = .01			α = .05		
	Fraction of data imputed			Fraction of data imputed		
n_1, n_2	0.100	0.300	0.500	0.100	0.300	0.500
10, 10	0.899	0.865	0.765	0.966	0.947	0.888
20, 20	0.992	0.982	0.953	0.995	0.998	0.992
30, 30	1.000	0.999	0.983	1.000	0.999	1.000
40, 40	1.000	1.000	1.000	1.000	1.000	1.000
60, 60	1.000	1.000	1.000	1.000	1.000	1.000
10, 30	0.976	0.955	0.923	0.991	0.992	0.981
20, 60	0.999	0.999	0.997	0.999	1.000	0.999
30, 90	1.000	1.000	1.000	1.000	1.000	1.000

Tables 16 – 20 present the number of rejections out of 1,000 for the *t* distribution under varying sample sizes, effect sizes, fractions of data imputed, and alpha. The *t* distribution has a single parameter: *df*. The distribution was simulated using *df* = 3.

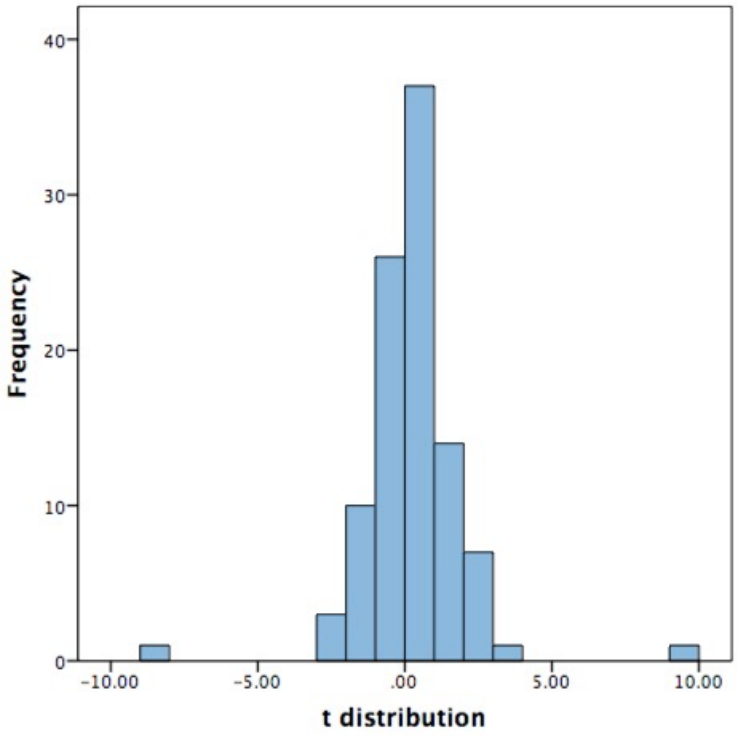


Figure 8. *t* distribution (*df* = 3). Created using SPSS syntax.

The probability density function for the *t* distribution (Figure 8) is

$$\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

Table 16

Small (0.2σ) Treatment Effect Rejection Rates for the t Distribution (df=3)

	α = .01			α = .05		
	Fraction of data imputed			Fraction of data imputed		
n_1, n_2	0.100	0.300	0.500	0.100	0.300	0.500
10, 10	0.026	0.024	0.021	0.098	0.086	0.078
20, 20	0.047	0.022	0.025	0.117	0.106	0.095
30, 30	0.026	0.029	0.035	0.139	0.114	0.123
40, 40	0.060	0.044	0.030	0.153	0.145	0.125
60, 60	0.059	0.076	0.063	0.211	0.183	0.156
10, 30	0.027	0.021	0.016	0.112	0.107	0.087
20, 60	0.035	0.041	0.035	0.137	0.118	0.115
30, 90	0.066	0.047	0.064	0.168	0.163	0.163

Table 17

Medium (0.5σ) Treatment Effect Rejection Rates for the t Distribution (df=3)

	α = .01			α = .05		
	Fraction of data imputed			Fraction of data imputed		
n_1, n_2	0.100	0.300	0.500	0.100	0.300	0.500
10, 10	0.138	0.125	0.090	0.311	0.290	0.211
20, 20	0.224	0.178	0.150	0.411	0.395	0.332
30, 30	0.318	0.268	0.202	0.543	0.515	0.421
40, 40	0.384	0.363	0.288	0.624	0.612	0.511
60, 60	0.575	0.508	0.402	0.760	0.725	0.640
10, 30	0.171	0.151	0.147	0.329	0.306	0.286
20, 60	0.289	0.290	0.237	0.491	0.500	0.469
30, 90	0.439	0.432	0.397	0.669	0.620	0.596

Table 18

Large (0.8σ) Treatment Effect Rejection Rates for the t Distribution (df=3)

	α = .01			α = .05		
	Fraction of data imputed			Fraction of data imputed		
n_1, n_2	0.100	0.300	0.500	0.100	0.300	0.500
10, 10	0.353	0.298	0.259	0.564	0.480	0.393
20, 20	0.555	0.514	0.397	0.749	0.698	0.599
30, 30	0.716	0.642	0.514	0.856	0.798	0.737
40, 40	0.805	0.749	0.610	0.919	0.887	0.813
60, 60	0.921	0.887	0.803	0.970	0.955	0.927
10, 30	0.455	0.430	0.368	0.624	0.602	0.567
20, 60	0.707	0.673	0.593	0.837	0.834	0.780
30, 90	0.881	0.822	0.772	0.954	0.910	0.924

Table 19

Very Large (1.2σ) Treatment Effect Rejection Rates for the t Distribution (df=3)

	α = .01			α = .05		
	Fraction of data imputed			Fraction of data imputed		
n_1, n_2	0.100	0.300	0.500	0.100	0.300	0.500
10, 10	0.643	0.545	0.450	0.798	0.740	0.657
20, 20	0.852	0.770	0.687	0.940	0.910	0.877
30, 30	0.935	0.908	0.838	0.979	0.973	0.940
40, 40	0.975	0.970	0.903	0.999	0.991	0.969
60, 60	0.999	0.997	0.978	1.000	1.000	0.997
10, 30	0.785	0.743	0.703	0.894	0.859	0.834
20, 60	0.956	0.945	0.904	0.985	0.983	0.969
30, 90	0.993	0.986	0.975	0.998	0.996	0.994

Table 20

Huge (2.0σ) Treatment Effect Rejection Rates for the t Distribution (df=3)

	α = .01			α = .05		
	Fraction of data imputed			Fraction of data imputed		
n_1, n_2	0.100	0.300	0.500	0.100	0.300	0.500
10, 10	0.905	0.871	0.796	0.949	0.929	0.911
20, 20	0.986	0.990	0.947	0.998	0.992	0.980
30, 30	0.998	0.996	0.988	1.000	1.000	0.998
40, 40	1.000	0.999	0.998	1.000	1.000	1.000
60, 60	1.000	1.000	1.000	1.000	1.000	1.000
10, 30	0.971	0.963	0.939	0.988	0.986	0.974
20, 60	1.000	0.998	0.997	1.000	0.998	0.997
30, 90	1.000	1.000	0.999	1.000	1.000	1.000

CHAPTER 5 DISCUSSION

Table 21 was constructed to serve as a benchmark to separate the effects of power theory from those induced by MI. It was created by drawing two samples (X_1, X_2) of sizes (n_1, n_2) from the Normal distribution, adding the stated treatment effect to X_2, conducting a two-tailed independent samples *t* test, and advancing a counter if the test was significant at the given alpha level. The procedure was repeated 1,000 times for each combination of factor levels. The results show the rejection rates that would be expected under normal power theory and demonstrate the power advantage of larger (balanced) samples, larger effect sizes, and a less conservative alpha level.

Table 21

Rejection Rates for the Nonimputed Normal Distribution

	α = .01					α = .05				
	Effect Size					Effect Size				
n_1, n_2	0.2σ	0.5σ	0.8σ	1.2σ	2.0σ	0.2σ	0.5σ	0.8σ	1.2σ	2.0σ
10, 10	0.018	0.049	0.171	0.440	0.927	0.059	0.190	0.385	0.725	0.987
20, 20	0.019	0.152	0.409	0.871	0.999	0.101	0.327	0.701	0.961	1.000
30, 30	0.023	0.255	0.680	0.972	1.000	0.123	0.478	0.850	0.995	1.000
40, 40	0.043	0.337	0.845	0.993	1.000	0.154	0.584	0.949	1.000	1.000
60, 60	0.069	0.563	0.960	1.000	1.000	0.168	0.774	0.994	1.000	1.000
10, 30	0.021	0.098	0.301	0.716	0.996	0.076	0.262	0.558	0.897	0.999
20, 60	0.023	0.264	0.641	0.971	1.000	0.129	0.486	0.860	0.991	1.000
30, 90	0.047	0.408	0.878	0.998	1.000	0.143	0.662	0.973	1.000	1.000

A small (0.2σ) difference between treatment arms was virtually undetectable (< 7%) at the .01 alpha level and marginally detectable (< 17%) at the .05 alpha level. Although clinicians would hope that a medium (0.5σ) treatment effect would be detected with a relatively small

sample, at the .01 alpha level, a sample size of 30, 30 yielded a one-in-four (.255) chance of rejection. At the .05 alpha level, clinicians had slightly less than a 50-50 (.478) chance of detecting a difference between treatment arms with the same 30, 30 sample size. Unbalanced designs were less powerful; a sample size of 20, 60 was necessary to achieve the same one-in-four (.264) results at alpha .01 and one-in-two (.486) results at alpha .05. A large (0.8σ) difference between treatment arms was detectable at Cohen's recommended .80 ($\beta = .20$) level once the sample size for balanced designs reached 40, 40 at alpha .01 and 30, 30 at alpha .05. Unbalanced designs reached the recommended power level once the sample size reached 30, 90 at alpha .01 and 20, 60 at alpha .05. A very large (1.2σ) treatment effect was detectable at the recommended .80 power level for all of the sample sizes save for the 10, 10 and 10, 30 levels at alpha .01 and the 10, 10 level at alpha .05. A huge difference between treatment arms was detectable at near maximum power for all of the sample sizes tested at both alpha levels.

The rejection rates presented in Tables 6 – 20 can be used to aid clinicians in determining the power that they can expect to achieve given a specific distribution, effect size, sample size, fraction of data imputed, and alpha. The results show that the rejection rates for imputed data follow the trends that would be expected under normal power theory regardless of distribution. There was a positive relationship between effect size and rejection rates with minimal power at the small (0.2σ) effect size level and near maximum power at the huge (2.0σ) effect size level. As expected, there was a positive relationship between sample size and rejection rates with balanced samples being more powerful than unbalanced samples. Power was also higher at the less conservative .05 alpha level.

Tables 22 – 26 illustrate the impact of distribution shape on rejection rates by presenting the range of rejection rates (the distribution with the highest rejection rate minus the distribution

with the lowest rejection rate) for the three distributions under investigation. The largest differences tended to occur at the smallest balanced (10, 10) and unbalanced (10, 30) sample size levels. The Normal distribution tended to reject at a lower rate than the Chi-square and *t* distributions at the medium (0.5σ) and large (0.8σ) effect size levels, but rejected at a higher rate at the very large (1.2σ) effect size level.

With a small (0.2σ) treatment effect, the largest difference at the .01 alpha level was 2.3 percentage points and occurred at the 30, 30 sample size level with 50% of the data imputed between the Normal (.023) and Chi-square (.046) distributions. The largest difference at alpha .05 was 5.5% and occurred at the 30, 90 sample size level with 10% of the data imputed between the Normal (.137) and Chi-square (.192) distributions.

With a medium (0.5σ) treatment effect, the largest difference at the .01 alpha level was 9.9 percentage points and occurred at the 40, 40 sample size level with 30% of the data imputed between the Normal (.264) and *t* (.363) distributions. The largest difference at the .05 alpha level was 11.8% and occurred at the 10, 10 sample size level with 10% of the data imputed between the Normal distribution (.193) and the *t* distribution (.311).

With a large (0.8σ) treatment effect, the largest difference at the .01 alpha level was 13.9% and occurred at the 10, 10 sample size level with 50% of the data imputed between the Normal (.120) and *t* (.259) distributions. At the .05 alpha level, the largest difference was 14.5% and occurred at the 10, 10 sample size level with 10% of the data imputed between the Normal distribution (.419) and the *t* distribution (.564). Excluding the 10, 10 and 20, 20 balanced sample size levels and the 10, 30 unbalanced sample size level reduced the largest difference at alpha .01 to 7.2% and the largest difference at alpha .05 to 6.9%.

Table 22

Range of Distribution Rejection Rates for a Small (0.2σ) Treatment Effect

	α = .01			α = .05		
	Percentage of data imputed			Percentage of data imputed		
n_1, n_2	10%	30%	50%	10%	30%	50%
10, 10	0.8	1.0	0.6	3.0	0.4	1.4
20, 20	1.9	1.1	1.0	1.4	1.3	1.1
30, 30	1.4	1.7	2.3	1.4	1.1	2.2
40, 40	1.8	1.0	0.4	1.2	0.6	0.9
60, 60	2.0	1.8	0.7	1.6	1.9	4.5
10, 30	1.4	0.6	0.9	4.5	2.3	1.1
20, 60	0.8	0.7	1.2	1.4	4.0	1.5
30, 90	1.8	0.9	1.9	5.5	1.1	3.0

Table 23

Range of Distribution Rejection Rates for a Medium (0.5σ) Treatment Effect

	α = .01			α = .05		
	Percentage of data imputed			Percentage of data imputed		
n_1, n_2	10%	30%	50%	10%	30%	50%
10, 10	5.9	5.8	2.9	11.8	11.1	5.0
20, 20	5.0	4.6	6.1	3.8	7.9	8.1
30, 30	8.7	4.7	6.3	11.4	8.9	8.5
40, 40	2.2	9.9	6.6	7.3	7.7	6.2
60, 60	4.7	6.7	5.2	0.9	0.1	4.1
10, 30	5.7	5.5	6.5	8.8	8.5	7.2
20, 60	2.9	5.5	8.3	5.6	5.9	8.3
30, 90	4.9	8.6	8.8	2.7	3.2	2.0

Table 24

Range of Distribution Rejection Rates for a Large (0.8σ) Treatment Effect

	α = .01			α = .05		
	Percentage of data imputed			Percentage of data imputed		
n_1, n_2	10%	30%	50%	10%	30%	50%
10, 10	11.8	12.3	13.9	14.5	12.9	11.3
20, 20	7.2	11.4	11.3	4.9	9.4	8.5
30, 30	3.2	7.2	6.8	1.7	0.6	6.9
40, 40	1.5	2.3	5.8	2.5	1.2	1.0
60, 60	2.9	1.1	1.5	2.1	1.1	1.8
10, 30	10.7	12.7	9.4	6.9	5.8	9.3
20, 60	3.5	4.1	5.8	2.7	0.8	2.5
30, 90	2.5	3.1	1.5	1.9	4.6	0.8

Table 25

Range of Distribution Rejection Rates for a Very Large (1.2σ) Treatment Effect

	α = .01			α = .05		
	Percentage of data imputed			Percentage of data imputed		
n_1, n_2	10%	30%	50%	10%	30%	50%
10, 10	12.7	12.0	14.3	7.8	6.5	10.2
20, 20	2.6	4.1	6.4	2.5	3.1	4.3
30, 30	3.3	2.6	2.0	1.7	1.3	1.8
40, 40	2.2	2.1	3.2	0.5	0.8	2.2
60, 60	0.1	0.5	1.5	0.0	0.1	0.3
10, 30	6.2	5.1	10.0	0.8	0.4	4.0
20, 60	3.8	1.6	3.5	0.9	0.9	0.1
30, 90	1.2	1.1	1.6	0.3	0.4	0.3

Table 26

Range of Distribution Rejection Rates for a Huge (2.0σ) Treatment Effect

	α = .01			α = .05		
	Percentage of data imputed			Percentage of data imputed		
n_1, n_2	10%	30%	50%	10%	30%	50%
10, 10	5.1	2.2	3.2	4.2	3.8	2.3
20, 20	1.3	1.4	3.2	0.5	0.8	1.9
30, 30	0.2	0.4	1.3	0.0	0.1	0.2
40, 40	0.0	0.1	0.2	0.0	0.0	0.0
60, 60	0.0	0.0	0.0	0.0	0.0	0.0
10, 30	2.5	4.2	4.7	1.2	1.4	2.4
20, 60	0.1	0.2	0.3	0.1	0.2	0.3
30, 90	0.0	0.0	0.1	0.0	0.0	0.0

With a very large (1.2σ) treatment effect, the largest difference at the .01 alpha level was 14.3% and occurred at the 10, 10 sample size level with 50% of the data imputed between the Normal distribution (.320) and the Chi-square distribution (.463). At the .05 alpha level, the largest difference was 10.2% and occurred at the 10, 10 sample size level with 50% of the data imputed between the Normal (.555) and *t* (.657) distributions. Excluding the 10, 10 and 20, 20 balanced sample size levels and the 10, 30 unbalanced sample size level reduced the largest difference at alpha .01 to 3.8% and the largest difference at alpha .05 to 2.2%.

With a huge (2.0σ) treatment effect, the largest difference at the .01 alpha level was 5.1% and occurred at the 10, 10 sample size level with 10% of the data imputed between the Chi-square (.899) and Normal (.950) distributions. At the .05 alpha level, the largest difference was 4.2% and occurred at the 10, 10 sample size level with 10% of the data imputed between the *t* (.949) and Normal (.991) distributions. Excluding the 10, 10 and 20, 20 balanced sample size

levels and the 10, 30 unbalanced sample size level virtually eliminated any differences between the distributions.

Tables 27 – 41 present the change in the rejection rate as a result of the change in the percentage of data imputed (PDI) for the given distribution, effect size, sample size and alpha. The results show that although statistical power decreased (or Type II error increased) as the PDI increased, the magnitude of the power loss was dependent upon effect size. As effect size increased, the change in rejection rate (as a function of the change in PDI) increased until the effect size reached the large (1.2σ) level at which point the trend reversed itself. Had the study not included Sawilowsky's (2009) very large and huge effect size levels, this finding would have been missed.

With a small (0.2σ) treatment effect (Tables 27 – 29), as the PDI increased from 10% to 30%, the loss of power was 2.5% or less at alpha .01 and 3.6% or less at alpha .05 regardless of sample size and distribution. As the PDI increased from 30% to 50%, the loss of power was 2.3% or less at alpha .01 and 4.4% or less at alpha .05 regardless of sample size and distribution. As the PDI increased from 10% to 50%, the loss of power was 3% or less at alpha .01 and 7.5% or less at alpha .05 regardless of sample size and distribution.

With a medium (0.5σ) treatment effect (Tables 30 – 32), as the PDI increased from 10% to 30%, the loss of power was 10% or less regardless of sample size for the Normal distribution and 6.7% or less regardless of sample size for the Chi-square and *t* distributions at alpha .01. At the .05 alpha level, the loss of power was 7% or less regardless of sample size and distribution. As the PDI increased from 30% to 50%, the loss of power was 12.6% or less regardless of sample size, distribution, and alpha. As the PDI increased from 10% to 50%, the loss of power was 17.8% or less regardless of sample size, distribution, and alpha.

Table 27

Impact of PDI on Normal Distribution, Small (0.2σ) Effect Size Rejection Rates

	α = .01			α = .05		
	% Change in Rejection Rate			% Change in Rejection Rate		
n_1, n_2	10-30%	30-50%	10-50%	10-30%	30-50%	10-50%
10, 10	0.3	-1.1	-0.8	1.4	-1.8	-0.4
20, 20	-0.5	-0.3	-0.8	0.1	-0.9	-0.8
30, 30	-0.4	-0.9	-1.3	-3.1	-0.6	-3.7
40, 40	0.1	-0.9	-0.8	-2.6	-2.3	-4.9
60, 60	-0.9	-0.2	-1.1	-3.1	-4.4	-7.5
10, 30	-1.7	0.6	-1.1	1.7	-0.8	0.9
20, 60	0.7	0.5	1.2	-2.2	-2.6	-4.8
30, 90	-0.3	0.5	0.2	1.5	-0.2	1.3

Table 28

Impact of PDI on X^2 Distribution, Small (0.2σ) Effect Size Rejection Rates

	α = .01			α = .05		
	% Change in Rejection Rate			% Change in Rejection Rate		
n_1, n_2	10-30%	30-50%	10-50%	10-30%	30-50%	10-50%
10, 10	-1.0	-0.7	-1.7	-0.8	-0.9	-1.7
20, 20	-0.6	-1.8	-2.4	1.2	-1.1	0.1
30, 30	0.6	0.0	0.6	-3.4	0.4	-3.0
40, 40	0.0	-2.3	-2.3	-2.1	-1.5	-3.6
60, 60	-1.2	-0.4	-1.6	-3.6	-0.2	-3.8
10, 30	0.3	-0.8	-0.5	-1.5	-0.5	-2.0
20, 60	0.4	-1.0	-0.6	2.4	-4.4	-2.0
30, 90	0.1	-0.9	-0.8	-3.0	-2.9	-5.9

Table 29

Impact of PDI on t Distribution, Small (0.2σ) Effect Size Rejection Rates

	α = .01			α = .05		
	% Change in Rejection Rate			% Change in Rejection Rate		
n_1, n_2	10-30%	30-50%	10-50%	10-30%	30-50%	10-50%
10, 10	-0.2	-0.3	-0.5	-1.2	-0.8	-2.0
20, 20	-2.5	0.3	-2.2	-1.1	-1.1	-2.2
30, 30	0.3	0.6	0.9	-2.5	0.9	-1.6
40, 40	-1.6	-1.4	-3.0	-0.8	-2.0	-2.8
60, 60	1.7	-1.3	0.4	-2.8	-2.7	-5.5
10, 30	-0.6	-0.5	-1.1	-0.5	-2.0	-2.5
20, 60	0.6	-0.6	0.0	-1.9	-0.3	-2.2
30, 90	-1.9	1.7	-0.2	-0.5	0.0	-0.5

Table 30

Impact of PDI on Normal Distribution, Medium (0.5σ) Effect Size Rejection Rates

	α = .01			α = .05		
	% Change in Rejection Rate			% Change in Rejection Rate		
n_1, n_2	10-30%	30-50%	10-50%	10-30%	30-50%	10-50%
10, 10	-1.8	-0.2	-2.0	-1.4	-0.7	-2.1
20, 20	-2.3	-4.4	-6.7	-5.9	-6.5	-12.4
30, 30	-1.0	-8.2	-9.2	-0.3	-9.0	-9.3
40, 40	-10.0	-4.2	-14.2	-2.3	-8.6	-10.9
60, 60	-8.7	-9.1	-17.8	-3.4	-12.6	-16.0
10, 30	0.4	-2.5	-2.1	0.1	-5.0	-4.9
20, 60	-3.1	-6.3	-9.4	-3.2	-6.3	-9.5
30, 90	-6.0	-3.7	-9.7	-6.6	-1.2	-7.8

Table 31

Impact of PDI on X^2 Distribution, Medium (0.5σ) Effect Size Rejection Rates

	α = .01			α = .05		
	% Change in Rejection Rate			% Change in Rejection Rate		
n_1, n_2	10-30%	30-50%	10-50%	10-30%	30-50%	10-50%
10, 10	-2.9	-2.1	-5.0	-2.8	-2.8	-5.6
20, 20	-1.7	-2.9	-4.6	-2.6	-7.1	-9.7
30, 30	-6.6	-5.6	-12.2	-6.1	-6.0	-12.1
40, 40	-2.5	-10.4	-12.9	-6.2	-10.6	-16.8
60, 60	-5.4	-12.1	-17.5	-4.4	-11.5	-15.9
10, 30	2.6	-1.5	1.1	-0.2	-6.3	-6.5
20, 60	-1.4	-2.6	-4.0	-2.9	-4.0	-6.9
30, 90	-5.4	-3.5	-8.9	-7.0	-3.1	-10.1

Table 32

Impact of PDI on t Distribution, Medium (0.5σ) Effect Size Rejection Rates

	α = .01			α = .05		
	% Change in Rejection Rate			% Change in Rejection Rate		
n_1, n_2	10-30%	30-50%	10-50%	10-30%	30-50%	10-50%
10, 10	-1.3	-3.5	-4.8	-2.1	-7.9	-10.0
20, 20	-4.6	-2.8	-7.4	-1.6	-6.3	-7.9
30, 30	-5.0	-6.6	-11.6	-2.8	-9.4	-12.2
40, 40	-2.1	-7.5	-9.6	-1.2	-10.1	-11.3
60, 60	-6.7	-10.6	-17.3	-3.5	-8.5	-12.0
10, 30	-2.0	-0.4	-2.4	-2.3	-2.0	-4.3
20, 60	0.1	-5.3	-5.2	0.9	-3.1	-2.2
30, 90	-0.7	-3.5	-4.2	-4.9	-2.4	-7.3

Table 33

Impact of PDI on Normal Distribution, Large (0.8σ) Effect Size Rejection Rates

	α = .01			α = .05		
	% Change in Rejection Rate			% Change in Rejection Rate		
n_1, n_2	10-30%	30-50%	10-50%	10-30%	30-50%	10-50%
10, 10	-5.2	-6.3	-11.5	-6.4	-6.5	-12.9
20, 20	-8.3	-11.6	-19.9	-9.6	-9.0	-18.6
30, 30	-11.4	-12.4	-23.8	-3.9	-13.2	-17.1
40, 40	-6.4	-15.5	-21.9	-5.7	-5.9	-11.6
60, 60	-5.6	-10.6	-16.2	-2.5	-2.8	-5.3
10, 30	-4.5	-2.2	-6.7	-2.0	-8.8	-10.8
20, 60	-3.5	-8.6	-12.1	-2.1	-7.1	-9.2
30, 90	-2.8	-8.5	-11.3	-1.0	-3.8	-4.8

Table 34

Impact of PDI on X^2 Distribution, Large (0.8σ) Effect Size Rejection Rates

	α = .01			α = .05		
	% Change in Rejection Rate			% Change in Rejection Rate		
n_1, n_2	10-30%	30-50%	10-50%	10-30%	30-50%	10-50%
10, 10	-1.9	-9.4	-11.3	-3.6	-8.1	-11.7
20, 20	-7.3	-10.5	-17.8	-4.8	-8.4	-13.2
30, 30	-8.5	-15.5	-24.0	-3.9	-9.9	-13.8
40, 40	-9.3	-8.3	-17.6	-2.0	-7.7	-9.7
60, 60	-3.9	-10.2	-14.1	-2.0	-3.7	-5.7
10, 30	-3.7	-3.6	-7.3	-3.1	-5.6	-8.7
20, 60	-0.5	-6.9	-7.4	-3.0	-3.8	-6.8
30, 90	-0.3	-8.4	-8.7	-2.2	-0.9	-3.1

Table 35

Impact of PDI on t Distribution, Large (0.8σ) Effect Size Rejection Rates

	α = .01			α = .05		
	% Change in Rejection Rate			% Change in Rejection Rate		
n_1, n_2	10-30%	30-50%	10-50%	10-30%	30-50%	10-50%
10, 10	-5.5	-3.9	-9.4	-8.4	-8.7	-17.1
20, 20	-4.1	-11.7	-15.8	-5.1	-9.9	-15.0
30, 30	-7.4	-12.8	-20.2	-5.8	-6.1	-11.9
40, 40	-5.6	-13.9	-19.5	-3.2	-7.4	-10.6
60, 60	-3.4	-8.4	-11.8	-1.5	-2.8	-4.3
10, 30	-2.5	-6.2	-8.7	-2.2	-3.5	-5.7
20, 60	-3.4	-8.0	-11.4	-0.3	-5.4	-5.7
30, 90	-5.9	-5.0	-10.9	-4.4	1.4	-3.0

Table 36

Impact of PDI on Normal Distribution, Very Large (1.2σ) Effect Size Rejection Rates

	α = .01			α = .05		
	% Change in Rejection Rate			% Change in Rejection Rate		
n_1, n_2	10-30%	30-50%	10-50%	10-30%	30-50%	10-50%
10, 10	-8.4	-11.2	-19.6	-4.5	-12.0	-16.5
20, 20	-6.7	-18.8	-25.5	-2.9	-9.1	-12.0
30, 30	-4.5	-8.2	-12.7	-1.2	-2.6	-3.8
40, 40	-1.2	-5.0	-6.2	-0.1	-0.7	-0.8
60, 60	-0.1	-0.6	-0.7	0.0	0.0	0.0
10, 30	-3.1	-8.9	-12.0	-4.0	-6.8	-10.8
20, 60	-1.1	-2.7	-3.8	-0.3	-2.3	-2.6
30, 90	-0.2	-0.6	-0.8	0.0	-0.3	-0.3

Table 37

Impact of PDI on X^2 Distribution, Very Large (1.2σ) Effect Size Rejection Rates

	α = .01			α = .05		
	% Change in Rejection Rate			% Change in Rejection Rate		
n_1, n_2	10-30%	30-50%	10-50%	10-30%	30-50%	10-50%
10, 10	-8.4	-8.9	-17.3	-4.6	-8.5	-13.1
20, 20	-6.3	-13.2	-19.5	-3.5	-5.3	-8.8
30, 30	-5.1	-7.6	-12.7	-1.3	-2.1	-3.4
40, 40	-2.7	-5.0	-7.7	-0.4	-0.6	-1.0
60, 60	-0.5	-1.4	-1.9	-0.1	0.0	-0.1
10, 30	-3.6	-6.0	-9.6	-4.2	-2.8	-7.0
20, 60	1.1	-4.6	-3.5	-0.5	-1.3	-1.8
30, 90	0.1	-1.0	-0.9	-0.1	-0.1	-0.2

Table 38

Impact of PDI on t Distribution, Very Large (1.2σ) Effect Size Rejection Rates

	α = .01			α = .05		
	% Change in Rejection Rate			% Change in Rejection Rate		
n_1, n_2	10-30%	30-50%	10-50%	10-30%	30-50%	10-50%
10, 10	-9.8	-9.5	-19.3	-5.8	-8.3	-14.1
20, 20	-8.2	-8.3	-16.5	-3.0	-3.3	-6.3
30, 30	-2.7	-7.0	-9.7	-0.6	-3.3	-3.9
40, 40	-0.5	-6.7	-7.2	-0.8	-2.2	-3.0
60, 60	-0.2	-1.9	-2.1	0.0	-0.3	-0.3
10, 30	-4.2	-4.0	-8.2	-3.5	-2.5	-6.0
20, 60	-1.1	-4.1	-5.2	-0.2	-1.4	-1.6
30, 90	-0.7	-1.1	-1.8	-0.2	-0.2	-0.4

Table 39

Impact of PDI on Normal Distribution, Huge (2.0σ) Effect Size Rejection Rates

	α = .01			α = .05		
	% Change in Rejection Rate			% Change in Rejection Rate		
n_1, n_2	10-30%	30-50%	10-50%	10-30%	30-50%	10-50%
10, 10	-6.3	-12.3	-18.6	-2.4	-6.3	-8.7
20, 20	-0.3	-1.7	-2.0	0.0	-0.1	-0.1
30, 30	0.0	-0.4	-0.4	0.0	0.0	0.0
40, 40	0.0	0.0	0.0	0.0	0.0	0.0
60, 60	0.0	0.0	0.0	0.0	0.0	0.0
10, 30	0.1	-2.7	-2.6	0.0	-0.2	-0.2
20, 60	0.0	0.0	0.0	0.0	0.0	0.0
30, 90	0.0	0.0	0.0	0.0	0.0	0.0

Table 40

Impact of PDI on X^2 Distribution, Huge (2.0σ) Effect Size Rejection Rates

	α = .01			α = .05		
	% Change in Rejection Rate			% Change in Rejection Rate		
n_1, n_2	10-30%	30-50%	10-50%	10-30%	30-50%	10-50%
10, 10	-3.4	-10.0	-13.4	-1.9	-5.9	-7.8
20, 20	-1.0	-2.9	-3.9	0.3	-0.6	-0.3
30, 30	-0.1	-1.6	-1.7	-0.1	0.1	0.0
40, 40	0.0	0.0	0.0	0.0	0.0	0.0
60, 60	0.0	0.0	0.0	0.0	0.0	0.0
10, 30	-2.1	-3.2	-5.3	0.1	-1.1	-1.0
20, 60	0.0	-0.2	-0.2	0.1	-0.1	0.0
30, 90	0.0	0.0	0.0	0.0	0.0	0.0

Table 41

Impact of PDI on t Distribution, Huge (2.0σ) Effect Size Rejection Rates

	α = .01			α = .05		
	% Change in Rejection Rate			% Change in Rejection Rate		
n_1, n_2	10-30%	30-50%	10-50%	10-30%	30-50%	10-50%
10, 10	-3.4	-7.5	-10.9	-2.0	-1.8	-3.8
20, 20	0.4	-4.3	-3.9	-0.6	-1.2	-1.8
30, 30	-0.2	-0.8	-1.0	0.0	-0.2	-0.2
40, 40	-0.1	-0.1	-0.2	0.0	0.0	0.0
60, 60	0.0	0.0	0.0	0.0	0.0	0.0
10, 30	-0.8	-2.4	-3.2	-0.2	-1.2	-1.4
20, 60	-0.2	-0.1	-0.3	-0.2	-0.1	-0.3
30, 90	0.0	-0.1	-0.1	0.0	0.0	0.0

With a large (0.8σ) treatment effect (Tables 33 – 35), as the PDI increased from 10% to 30%, the loss of power was 11.4% or less for the Normal and Chi-square distributions and 7.4% or less for the *t* distribution at alpha .01 regardless of sample size. At the .05 alpha level, the loss of power was 5.8% or less save for the 10, 10 and 20, 20 sample size levels regardless of distribution. As the PDI increased from 30% to 50%, the loss of power at alpha .01 was 15.5% or less regardless of sample size and distribution. At the .05 alpha level, the loss of power was 13.2% or less for the Normal distribution and 9.9% or less for the Chi-square and *t* distributions regardless of sample size. As the PDI increased from 10% to 50%, the loss of power was 24% or less at alpha .01 and 18.6% or less at alpha .05 regardless of sample size and distribution.

With a very large (1.2σ) treatment effect (Tables 36 – 38), as the PDI increased from 10% to 30%, the loss of power was 5.1% or less save for the 10, 10 and 20, 20 sample size levels at alpha .01 regardless of distribution. At the .05 alpha level, the loss of power was 1.3% or less

save for the 10, 10 and 20, 20 balanced sample size levels and the 10, 30 unbalanced sample size level regardless of distribution. As the PDI increased from 30% to 50%, the loss of power was 8.9% or less save for the 10, 10 and 20, 20 sample size levels at alpha .01 regardless of distribution. At alpha .05, the loss of power was 3.3% or less save for the 10, 10 and 20, 20 balanced sample size levels and the 10, 30 unbalanced sample size level regardless of distribution. As the PDI increased from 10% to 50%, the loss of power was 12.7% or less at alpha .01 regardless of sample size and distribution. At the .05 alpha level, the loss of power was 3.9% or less save for the 10, 10 and 20, 20 balanced sample size levels and the 10, 30 unbalanced sample size level regardless of distribution.

With a huge (2.0σ) treatment effect (Tables 39 – 41), the PDI had virtually no impact on rejection rates save for the 10, 10 and 20, 20 balanced sample size levels and the 10, 30 unbalanced sample size level regardless of distribution and alpha.

In conclusion, one of the strengths of multiple imputation is that it has power advantages over traditional missing data methods (see Chapter 2). As the results from this simulation indicate that there can be a loss of power when MI is utilized using the defaults in SPSS (m = 5, iterations = 10) and a limited imputation model (one auxiliary variable with r = .50), it is recommended that clinicians plan for the treatment of missing data in the design stage to reduce the probability of making a Type II error.

Limitations of the Study

Some degree of caution is warranted when generalizing the results of simulation studies to a broad range of settings. This study was conducted using SPSS; the use of other software could potentially affect results (e.g., Allison, 2000; Von Hippel, 2004). Imputations were generated using 10 iterations and m = 5. Changing either (or both) of these simulation

characteristics could affect rejection rates. One auxiliary variable (r = .50) was included in the imputation model, adding additional variables or changing the correlation of the variable in the model (see chapter 2) could affect results. Any generalizations should be limited to the factor levels simulated in this study.

Scope for Future Research

There are two key ways that this research could be extended. First, this Monte Carlo simulation investigated three theoretical distributions. As public access to clinical trial data increases, clinicians could improve the ecological validity of this research by estimating real clinical distributions. Second, this research was conducted under the MCAR assumption. It could be extended by examining missing data created under the MNAR mechanism. Of special interest would be an investigation into MI's limitations when the missing data follow an exceptionally complex pattern.

REFERENCES

Acock, A. C. (2005). Working with missing values. *Journal of Marriage and Family, 67,* 1012-1028.

Allison, P. D. (2000). Multiple imputation for missing data: A cautionary tale. *Sociological Methods and Research, 28*, 301-309.

Allison, P. D. (2001). *Missing data*. Thousand Oaks, CA: Sage.

Azar, B. (2002). Finding a solution for missing data. *Monitor on Psychology, 33*, 70.

Baraldi, A. N., & Enders, C. K. (2010). An introduction to modern missing data analysis. *Journal of School Psychology, 48,* 5-37. doi:10.1016/j.jsp.2009.10.001

Barnard, J., & Rubin. D. B. (1999). Small sample degrees of freedom with multiple imputation. *Biometrika*, 86(4), 948-955.

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

Barzi, F., & Woodward, M. (2004). Imputations of missing values in practice: Results from imputations of serum cholesterol in 28 cohort studies. *American Journal of Epidemiology*, 160(1), 34-45.

Berlin, T. R. (2009). Missing data: what a little can do, and what researchers can do in response. *American Journal of Ophthalmology, 148*(6), 820-822.

Bodner, T. E. (2008). What improves with increased missing data imputations? *Structural Equation Modeling: A Multidisciplinary Journal, 15,* 651-675.

Boneau, C. A. (1960). The effects of violations of assumptions underlying the *t* test. *Psychological Bulletin, 57*(1), 49-64.

Bradley, J. V. (1977). A common situation conductive to bizzare distribution shapes. *The American Statistician, 31*, 147-150.

Bridge, P. D., & Sawilowsky (1999). Increasing physicians' awareness of the impact of statistics on research outcomes: comparative power of the t-test and Wilcoxon Rank-Sum test in small sample research. *Journal of Clinical Epidemiology, 52*, 229-236. DOI:10.1016/S0895-4356(98)00168-1

Brown, C. G., Kelon, G., D., Ashton, J., J., & Werman, H. A. (1987). The beta error and sample size determination in clinical trials in emergency medicine. *Annals of Emergency Medicine, 16*(2), 183-187.

Buhi, E. R., Goodson, P., & Neilands, T. B. (2008). Out of sight, not out of mind: Strategies for handling missing data. *American Journal of Health Behavior, 32*(1), 83-92.

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

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, J. F., Robinson, E. S. J., & Munafo, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews of Neuroscience, 14*(5), 365-376. doi:10.1038/nm3475

Carpenter, J. R., Kenward, M. G., & White, I. R. (2007). Sensitivity analysis after multiple imputation under missing at random: a weighting approach. *Statistical Methods in Medical Research, 16*(3), 259-275.

Chase, L. J., & Chase, R. B. (1976). A statistical power analysis of applied psychological research. *Journal of Applied Psychology, 61*, 234-237.

Choi, Y., Golder, S., Gillmore, M. R., Morrison, D. M. (2005). Analysis with Missing Data in Social Work Research. *Journal of Social Service Research, 31*(3), 23-48.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology,* 65, 145-153.

Cohen, J. (1965). Some statistical issues in psychological research. In B. B. Wolman (Ed.), Handbook of clinical psychology, 95-121. New York: McGraw-Hill.

Cohen, J. (1988). *Statistical power for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45*(12) 1304-1312.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155-159.

Collins, L. M., Schafer, J. L., & Kam, C. –M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods, 6(*4), 330-351.

Davey, A., & Savla, J. (2010). *Statistical power analysis with missing data: A structural equation modeling approach*. New York: Routledge.

Demirtas, H., Freels, S. A., Yucel, R. M. (2008). Plausibility of multivariate normality assumption when multiply imputing non-Gaussian continuous outcomes: A simulation assessment. *Journal of Statistical Computation and Simulation, 78*(1), 69-84.

Edlund, M. J., Overall, J. E., Rhoades, H., M. (1985). Beta, or Type II error in psychiatric controlled clinical trials. *Journal of Psychiatric Research, 19*(4), 563-567.

Enders, C. K. (2001). The impact of nonnormality on full information maximum-likelihood estimation for structural equation models with missing data. *Psychological Methods, 6*(4), 352-370. doi:10.1037//1082-989X.6.4.352

Enders, C. K. (2006). A primer on the use of modern missing-data methods in psychosomatic medicine research. *Psychosomatic Medicine, 68,* 427-436.

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

Enders, 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: A Multidisciplinary Journal, 8*, 430-457.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*, 175-191.

Fichman, M., & Cummings, J. N. (2003). Multiple imputation for missing data: Makin the most of what you know. *Organizational Research Methods*, *6*(3), 282-308. doi:10.1177/1094428103255532

Fleishman, A. I. (1978). A method for simulating non-normal distributions. *Psychometrika, 43*(4), 521–532.

Freiman, J. A., Thomas, A. B., Chalmers, C., Smith, H., & Kuebler, R. R. (1978). The importance of beta, the type II error and sample size in the design and interpretation of the randomized controlled trial. *The New England Journal of Medicine, 299*, 690-694.

Grace, T. A., & Sawilowsky, S. S. (2009). Data error prevention and cleansing: A comprehensive guide for instructors of statistics and their students. *Model Assisted Statistics and Applications, 4*, 303-312. doi:10.3233/MAS-2009-0140

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology, 60,* 549-576. doi:10.1146/annurev.psych.58.110405.085530

Graham, J. W. (2012). *Missing data: Analysis and design*. New York: Springer.

Graham, J. W., & Donaldson, S. I. (1993). Evaluating interventions with differential attrition: The importance of nonresponse mechanisms and use of follow-up data. *Journal of Applied Psychology*, 78, 119-128.

Graham, J. W., Hofer, S. M., Donaldson, S. I., MacKinnon, D. P., & Schafer, J. L. (1997). Analysis with missing data in prevention research. In K. J. Bryant, M. Windle, & S. G. West (Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research* (pp. 335-353). Washington, DC: American Psychological Association.

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.

Graham, J. W., & Schafer, J. L. (1999). On the performance of multiple imputation for multivariate data with small sample size. In R. Hoyle (Ed.). *Statistical strategies for small sample research.* Thousand Oaks, CA: Sage.

Graham, J. W., & Schafer, J. L. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*, 147-177.

Hardt, J., Herke, M., & Leonhart, R. (2012). Auxiliary variables in multiple imputation in regression with missing X: a warning against including too many in small sample research. *BMC Medical Research Methodology*, *12*. doi:10.1186/1471-2288-12-184

Harel, O., Pellowski, J., & Kalichman, S. (2012). Are we missing the importance of missing values in HIV prevention randomized clinical trials? Reviews and recommendations. *AIDS Behavior, 16*(6), 1382-1393. doi:10.1007/s10461-011-0125-6

He, Y., & Raghunathan, T. E. (2009). On the performance of sequential regression multiple imputation methods with non normal error distributions. *Communications in Statistics-Simulation and Computation, 38*(4), 856-883.

Headrick, T. C., & Sawilowsky, S. S. (2000). Weighted simplex procedures for determining the boundary points and constants for the univariate and multivariate power methods. Journal of Educational and Behavioral Statistics, 25(4), 417-436.

Horton, N. J., Lipsitz, S. R., & Parzen, M. (2003). A potential for bias when rounding in multiple imputation. *The American Statistician, 57*(4), 229-232.

Howard, G. S., Maxwell, S. E., & Fleming, K. J. (2000). The proof of the pudding: An illustration of the relative strengths of null hypothesis, meta-analysis, and bayesian analysis. *Psychological Methods, 5*(3), 315-332.

Huber, P. J. (1972). Robust statistics: A review. *Annals of Mathematical Statistics, 43*, 1041-1067.

IBM Corp. (2011). *IBM SPSS (20) Algorithms*. NY: IBM Corp.

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

Kearon, C., Ginsberg, J. S., Julian, J. A., Douketis, J., Solymoss, S., Ockelford, P., et al. (2006). Comparison of fixed dose weight adjusted unfractionaed heparin and low molecular weight heparin for acute treatment of venous thromboembolism. *Journal of the American Medical Association, 296*(8), 935-942.

Kim, J. (2004). Finite sample properties of multiple imputation estimators. *The Annals of Statistics, 32*(2), 766-783. doi:10.1214/009053604000000175

Klebanoff, M. A., & Cole, S. R. (2008). Use of multiple imputation in the epidemiologic literature. *American Journal of Epidemiology, 168*(4), 355-357. doi:10.1093/aje/kwn071

Lee, K. J. & Carlin, J. B. (2012). Recovery of information from multiple imputation: a simulation study. *Emerging Themes in Epidemiology, 9*(3), 1-10. doi:10.1186/1742-7622-9-3

Levene, H. (1960). Robust tests for equality of variances. In I. Olkin (Ed.), *Contributions to probability and statistics: Essays in honor of Harold Hotelling (pp. 278-292).* Palo Alto, CA: Stanford University Press.

Little, R. J. A. (1988). Robust estimation of the mean and covariate matrix from data with missing values. *Applied Statistics, 37,* 23-38.

Little, R. J. A. (1992). Regression with missing X's: A review. *Journal of the American Statistical Association, 90,* 1112-1121.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data*. New York: Wiley.

Littell, R. C., Roderick, J. A., & Schenker, N. (1995). Missing data. In G. Arminger, C. Clogg, & M. Sobel (Eds.). *Handbook of statistical modeling for the social and behavioral sciences* (pp. 39-75). New York: Plenum.

Lix, L. M., Keselman, J. C., & Keselman, H. J. (1996). Consequences of assumption violations revisited: A quantitative review of alternatives to the one-way analysis of variance "F" test. *Review of Educational Research, 66,* 579-619.

Mackinnon, A. (2010). The use and reporting of multiple imputation in medical research – a review. *Journal of Internal Medicine, 268,* 586-593. doi: 10.1111/j.1365-2796.2010.02274.x

Marshall, A., Altman, D. G., Royston, P., & Holder, R. L. (2010). Comparison of techniques for handling missing covariate data within prognostic modeling studies: A simulation study. *BMC Medical Research Methodology, 10.*

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: causes, consequences, and remedies. *Psychological Methods, 9*(2), 147-163.

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

Micceri. T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105*(1), 156-166.

Moher, D., Dulberg, C. S., & Wells, G. A. (1994). Statistical power, sample size, and their reporting in randomized controlled trials. *Journal of the American Medical Association*, 272(2), 122-124. doi:10.1001/jama.1994.03520020048013

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

Murphy, K. R., Myors, B., & Wolach, A. (2009). *Statistical power analysis*. New York, New York: Routledge.

National Academy of Science. (2010). *The prevention and treatment of missing data in clinical trials.* Panel on Handling Missing Data in Clinical Trials. Committee on National Statistics, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

Olinsky, A., Chen, S. Y., & Harlow, L. (2003). The comparative efficacy of imputation methods for missing data in structural equation modeling. *European Journal of Operational Research, 15,* 53-79.

O'Neill, R. T., & Temple, R. (2012). The prevention and treatment of missing data in clinical trials: An FDA perspective on the importance of dealing with it. *Clinical Pharmacology Therapeutics, 91*, 550-554.

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

Pigott, T. D. (2001). A review of methods for missing data. *Educational Research and Evaluation: An International Journal on Theory and Practice, 7*(4), 353-383.

Pike, J., & Leith, J. (2009). Type II error in the shoulder and elbow literature. *Journal of Shoulder and Elbow Surgery* 18, 44 – 51.

Raghunathan, T. E. (2004). What do we do with missing data? Some options for analysis of incomplete data. *Annual Review of Public Health, 25*, 99-117.

Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J., & Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology, 27*(1), 85-95.

Rässler, S., Rubin, D. B., & Zell, E. R. (2013). Imputation. *Computational Statistics*, 5, 20-29. doi:10.1002/wics.1240

Rosenthal, R. (1990). How are we doing in soft psychology? *American Psychologist, 45*, 775-777.

Rossi, J. S. (1990). Statistical power of psychological research: What have we gained in 20 years? *Journal of Consulting and Clinical Psychology, 58,* 646-656.

Rubin, D. B. (1976) Inference and missing data. *Biometrika, 63*(3), 581-590.

Rubin, D. B. (1987). Multiple imputation for nonresponse in surveys. New York: Wiley.

Rubin, D. B. (1996). Multiple imputation after 18+ years (with discussion). *Journal of the American Statistical Association, 91*, 473-489.

Sawilowsky, S. S. (2009). New effect size rules of thumb, *Journal of Modern Applied Statistical Methods (8)*2, 597-599.

Sawilowsky, S. S., & Blair, R. C. (1992). A more realistic look at the robustness and Type II error properties of the t-test to departures from population normality. *Psychological Bulletin*, *111*(2), 353-360.

Sawilowsky, S. S., & Hillman, S. B. (1992). Power of the independent samples t test under a prevalent psychometric measure distribution. *Journal of Consulting and Clinical Psychology, 60*(2), 240-243.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. Boca Raton, FL: Chapman & Hall.

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

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*(2), 147-177.

Schafer, J. L., & Olsen, M. K. (1998). Multiple imputation for multivariate missing data problems: A data analyst's perspective. *Multivariate Behavioral Research, 33*, 545-571.

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. doi:10.1037/a00118082

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.

Sterne, J. A. C., White, I. R., Carlin, J. B., Spratt, M. Royston, P., Kenward, M. G., Wood, A. M., Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: potential pitfalls. *British Medical Journal, 338*. doi:10.1136/bmj.b2393

Stuart, E. A., Azur, M., Frangakis, C. & Leaf, P. (2009). Multiple imputation with large data sets: A case study of the children's mental health initiative. *American Journal of Epidemiology,* 169(9), 1133-1139. doi:10.1093/aje/kwp026

Student (1908). The probable error of a mean. *Biometrika, 6*(1), 1-25.

Tabachnick, B. G., & Fiddell, L. S. (2007). *Using Multivariate Statistics* (5th ed.). Boston: Pearson/Allyn & Bacon.

Tsang, R., Colley, L., & Lynd, L. D. (2009). Inadequate statistical power to detect clinically significant differences in adverse event rates in randomized controlled trials. *Journal of Clinical Epidemiology, 62*, 609-616.

Van Buuren, S. (2012). *Flexible imputation of missing data*. Boca Raton, FL: CRC Press

Van Buuren, S., Brand, J. P. L., Groothuis-Oudshoorn, C. G. M. & Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation, 76*(12), 1049-1064.

Van Buuren, S., & Groothuis-Oudshoorn, C. G. M. (1999). *Flexible multivariate imputation by MICE*. Technical Report PG/VGZ/99.054, TNO Prevention and Health, Leiden.

Von Hippel, P. T. (2004). Biases in SPSS 12.0 missing value analysis. *The American Statistician, 58*(2), 160-164. doi:10.1198/0003130043204

Von Hippel, P. T. (2005). How many imputations are needed? A comment on Hershberger and Fisher (2003). *Structural Equation Modeling,* 12, 334-335.

Von Hippel, P. T. (2013a). Should a normal imputation model be modified to impute skewed variables? *Sociological Methods and Research, 42*(1), 105-138.

Von Hippel, P. T. (2013b). The bias and efficiency of incomplete-data estimators in small univariate normal samples. *Sociological Methods & Research, 42*(4), 531-588.

White, I. R., Royston, P., & Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine, 30*, 377-399. doi:10.1002/sim.4067

Wilcox, R. R. (1998). How many discoveries have been lost by ignoring modern statistical methods? *American Psychologist, 53*, 300-314.

Wilkinson, L., & Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*, 594-604.

Williams, H. C., & Seed, P. (1992). Inadequate size of 'negative' clinical trials in dermatology. *British Journal of Dermatology, 128*, 317-326.

Wood, A. M., White, R. R., & Thompson, S. G. (2004). Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals. *Clinical Trials, 1*, 368-376.

Woods, S. P., Rippeth, J. D., Conover, E., Carey, C. L., Parsons, T. D., Tröster, A. I. (2006). Statistical power of studies examining the cognitive effects of subthalmic nucleus deep brain stimulation in parkinson's disease. *The Clinical Neuropsychologist, 20*, 27-38. doi:10.1080/13854040500203290

Young, W., Weckman, G., & Holland, W. (2011). A survey of methodologies for the treatment of missing values within datasets: Limitations and benefits. Theoretical Issues in Ergonomics Science, 12(1), 15-43. doi:10.1080/14639220903470205

Yuan, K. H., & Bentler, P. M. (1999). On normal theory and associated test statistics in covariance structure analysis under two classes of nonnormal distributions. *Statistica Sinica, 9*, 831-853.

Yuen, S. Y., & Pope, J. E. (2008). Learning from past mistakes: assessing trial quality, power and eligibility in non-renal systemic lupus erythematosus randomized controlled trials. Rheumatology, 47, 1367-1372. doi:10.1093/rheumatology/ken230

Zimmerman, D. W. (1987). Comparative Power of Student t test and Mann-Whitney U test for unequal sample sizes and variances. *Journal of Experimental Education, 55*(3), 171-174.

ABSTRACT

THE IMPACT OF MULTIPLE IMPUTATION ON THE TYPE II ERROR RATE OF THE T TEST

by

TAMMY A. GRACE

August 2016

Advisor: Shlomo Sawilowsky, PhD

Major: Education Evaluation and Research

Degree: Doctor of Philosophy

The National Academy of Science identified numerous high priority areas for missing data research. This study addresses several of those areas by systematically investigating the impact of multiple imputation on the rejection rate of the independent samples *t* test under varying conditions of sample size, effect size, fraction of missing data, distribution shape, and alpha. In addition to addressing gaps in the missing data literature, this study also provides an overview of the multiple imputation procedure, as implemented in SPSS, with a focus on the practical aspects and challenges of using this method.

AUTOBIOGRAPHICAL STATEMENT

TAMMY A. GRACE

EDUCATION

2016	Doctor of Philosophy Wayne State University, Detroit, Michigan Dissertation defended June, 2016 Major: Education Evaluation and Research Specialization: Quantitative Research Cognate: Nonprofit Management
2002	Masters of Arts University of Michigan, Dearborn, Michigan (in collaboration with the Institute for Social Research, Ann Arbor, MI) Major: Liberal Studies Specialization: Research Methodology
1993	Bachelor of Arts University of Michigan, Dearborn, Michigan Majors: Psychology and Sociology

PUBLICATIONS

Grace, T. & Sawilowsky, S. (2009). Data error prevention and cleansing: A comprehensive guide for instructors of statistics and their students. *Model Assisted Statistics and Applications 4*, 303-312.