Analytical Methods for Bias Reduction: Cluster Randomized Trials with Longitudinal Data

By

Wairimu Magua

A dissertation submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

(Industrial and Systems Engineering)

at the

UNIVERSITY OF WISCONSIN – MADISON

2014

Date of Final Examination: 8/12/14

The dissertation is approved by the following members of the Final Oral Committee:

Mary L. (Molly) Carnes, Professor, Medicine & Industrial and Systems Engineering
Mari Palta, Professor, Population Health Sciences & Biostatistics & Medical Informatics
John D. Lee, Professor, Industrial and Systems Engineering
Laura A. McLay, Associate Professor, Industrial and Systems Engineering
Angela Byars-Winston, Professor, Medicine
Douglass L. Henderson, Professor, Engineering Physics

ACKNOWLEGEMENTS

This dissertation represents the culmination of a great learning experience here at the University of Wisconsin-Madison. I feel very privileged to be a part of the UW community. First, I would like to express my deep gratitude to my adviser and chair of the doctoral committee, Professor Molly Carnes, for not only giving me the great opportunity to work on this research but for also being a great mentor and role model as I navigated the academic labyrinth. I am also deeply grateful to Professor Mari Palta for not only being an uncompromisingly great educator but also for her generous guidance and mentorship throughout this and other projects. I thank Professor Angela Byars-Winston for her academic advice and guidance through difficult periods. I have had the privilege of being a part of a wonderful team at the Center for Women's Health Research for the past three years. The support and encouragement that I received has been invaluable. I thank Professor David Zimmerman for his guidance in my early part of the graduate school and Professor Douglass Henderson for being a generous mentor and for consistently engaging me in discussions about my academic progress. Also, I thank Professor John Lee for his guidance in my doctoral work and Professor Laura McLay for agreeing to be a part of my committee on very short notice.

I am very grateful to my parents for anchoring my confidence and my persistence in this journey with unwavering love and support. I would also like to thank my siblings for being a great source of support and laughter, especially in the most difficult of moments.

I have also had the great privilege of having wonderful friends. Each one has been, in their own special way, a source of refuge, reflection, and support. My acknowledgement would be incomplete without expressing deep love and appreciation to my husband Aaron for reminding me of my dreams and for his uncompromising support. He has continually and selflessly nurtured my aspirations. I extend my deep gratitude to my newborn son Samuel. He has been a source of daily joy as I worked on my dissertation, and he will continue to be hereafter.

Table of Contents

ABSTRACT

(Word Count = 350)

A cluster randomized trial (CRT) is a randomized controlled trial in which the units of randomization are aggregates of individuals known as clusters (A. Donner & Klar, 2002; Higgins & Green, 2008). Such designs are particularly appropriate for evaluating interventions that are implemented on the community, clinic, or hospital level. CRTs can be used to investigate costly quality improvement (QI) interventions with numerous outcomes when the best approach to achieve improvement is uncertain (Gustafson et al., 2013; Samsa & Matchar, 2000).

Randomization within cluster strata has been widely used to mitigate imbalance in baseline covariates. However, biased evaluation can still arise due to non-participation of individuals. CRTs with a longitudinal component are especially at risk for differential attrition (S. Eldridge, Ashby, Bennett, Wakelin, & Feder, 2008). Propensity scores have been widely used in the analysis of data from observational studies as an adjustment method to reduce selection bias (Paul R. Rosenbaum & Rubin, 1983). It was not until recently that propensity score based methods were applied to a CRT as a means of mitigating selection bias (Leyrat, Caille, Donner, & Giraudeau, 2013). Methods to address bias from differential attrition, a likely scenario in CRTs with longitudinal component, have not been explored.

The objective of this research was to use propensity score methods to address selection bias and bias arising from differential attrition in CRTs with longitudinal data. Using data from a CRT analyzed by mixed effects models, propensity scores were used to adjust for covariate imbalance and differential attrition.

This research confirms that missingness depending on random effects can be captured by missingness patterns and the number of observations (Park, Palta, Shao, & Shen, 2002). Furthermore, propensity scores estimated from the number of observations and missing data patterns were moderately successful at bias reduction. The research findings suggest that applying propensity scores as an interaction term rather than a covariate term is more effective in reducing bias.

This study establishes that propensity score based methods can be used to address bias from differential attrition in CRT's with a longitudinal component. It contributes pragmatic ways of addressing bias in such trials.

CHAPTER 1
INTRODUCTION

1.1. Overview

Randomized controlled trials (RCTs) and cluster randomized trials (CRTs) are acknowledged as standard quantitative evaluative designs for quality improvement research in healthcare (Eccles, Grimshaw, Campbell, & Ramsay, 2003). The agency for healthcare research and quality (AHRQ), a leading United States Department of Health and Human Services agency that funds interdisciplinary research from various disciplines including industrial engineering, organizational theory, and human factors, among others, has supported research investigating the role of CRTs in evaluating quality improvement interventions (Garrison & Mangione-Smith, 2013; Loeb, 2002; Kathleen N. Lohr, 2007; Mazor et al., 2007).

A RCT is considered to be the most rigorous experimental design for testing whether a causal relationship exists between an intervention and an outcome (Bonnie & Martin, 1998). One of most salient features of the RCT is random allocation of subjects to the intervention and control groups because it improves the likelihood of having comparable groups before the intervention. Cluster randomized trials, also known as group randomized trials, are a type of RCT in which groups of individuals referred to as clusters are randomized (M. K. Campbell, Elbourne, & Altman, 2004; Cornfield, 1978; A. Donner & Klar, 2002; Higgins & Green, 2008; Murrary, 2002). Types of clusters in

health systems research include hospitals, clinics, academic institutions, units within institutions, and geographically defined communities, among others. Unlike individual randomized trials, the unit of randomization (cluster) and the unit of analysis (cluster or individual) in CRTs are not necessarily the same.

In CRTs, individuals within a cluster tend to respond more similarly than individuals in different clusters. Hence, individual level data cannot be assumed to be independent. As a result, CRTs are considered to be less efficient (Walsh, 1947) than the individual randomized trials because sample sizes need to account for the within-cluster dependency. Still, the use of random allocation and control groups makes a CRT an important experimental design when evaluating interventions that cannot be delivered at the individual level. CRTs, by design, can minimize group contamination or contamination bias especially when compared to unblinded individual randomized controlled trials. CRTs can also be used to enhance compliance with an intervention. In the case of CRT with random allocation at a provider level, for example, patients clustered under the same provider may share information hence enhancing compliance (Glynn, Brookhart, Stedman, Avorn, & Solomon, 2007).

This chapter focuses on the importance of quality improvement (QI) in transforming healthcare system to a synchronized delivery system and the role of CRTs in evaluating the impact of QI interventions. While methodological developments of CRTs have lagged behind those of individually randomized trials, CRTs can provide opportunities to rigorously test the effectiveness of a wide range of interventions in the healthcare system. The objective of this research is to develop propensity score methods in CRTs with a longitudinal component to address specific types of biases that arise from

differential participation leading to imbalance in measured and unmeasured covariates. This study represents practical and vital contributions towards advancing CRT analytic approaches to handle the evaluation of complex quality improvement interventions in healthcare.

1.2. Quality Improvement and the Role of Cluster Randomized Trials

Many challenges in the healthcare system have been identified by the Institute of Medicine (IOM) indicating that potential interventions are better implemented on levels that involve aggregates of patients or practitioners such as clinics or hospitals rather than at individual levels (M. Campbell, Fitzpatrick, Haines, Kinmonth, & et al., 2000; M. K. Campbell, et al., 2004). The health care system is plagued by suboptimal quality, care discontinuity, high costs and inadequate access (Betancourt & Maina, 2004; Chassin & Galvin, 1998; Kohn, Corrigan, & Donaldson, 2000; Landrigan et al., 2010; McGlynn et al., 2003). In 2005, with the aim of transforming the over $1.6 trillion spending in the health sector, the National Academy of Engineering (NAE) and Institute of Medicine (IOM), set forth a framework for a systems approach anchored in an engineering and healthcare professional partnership (Reid et al., 2005). The objective of the partnership is to transform the U.S healthcare system, a conglomerate of loosely connected entities (laboratories, nursing homes, clinics, hospitals, pharmacies, community health centers, etc.), to a synchronized healthcare delivery system that meets six quality aims: safety, effectiveness, patient centeredness, timeliness, efficiency, and equitability. The safety component requires the integration of patient experiences and perspectives in all aspects of safety initiatives, hence extending safety beyond healthcare providers and institutions (P. F. Brennan & Safran, 2004). The

relevance of the NAE and IOM partnership remains salient given that healthcare spending by federal and local governments in the U.S. is projected to be $2.4 trillion (49% of national health spending) by 2022 (Cuckler et al., 2013).

The IOM model of healthcare was adopted from Ferlie and Shortell (Ferlie & Shortell, 2001). This model has four nested levels: patients; care teams; organizations; and the environment making the cluster randomized trial ideal for studying quality improvement interventions (QI). The UK Medical Research Council (MRC) defines a complex system as one with interacting multi-dimensional components (Craig et al., 2008a) and provides guidance to the development and evaluation of randomized trial interventions in complex systems such as healthcare (M. Campbell, et al., 2000; N. C. Campbell et al., 2007). The MRC framework can be used to guide the development of QI interventions

in the different levels of the IOM model of healthcare (Reid, et al., 2005) (Figure 1).

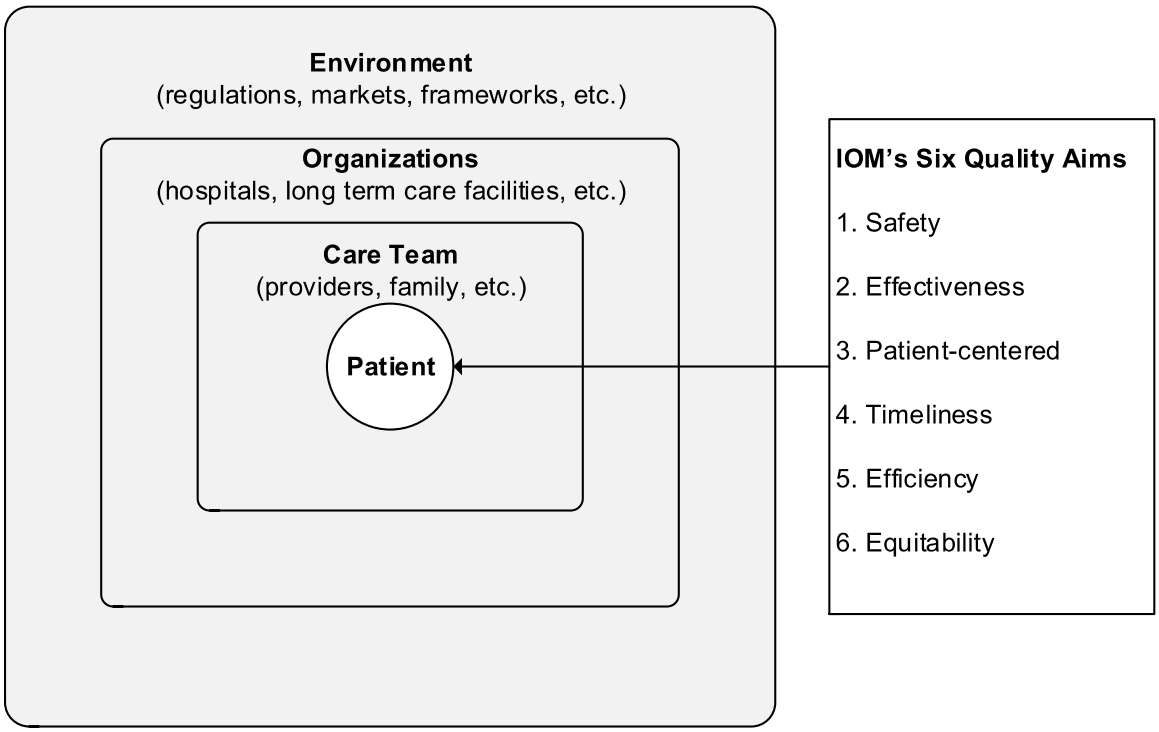

Figure 1. CRTs can be applied to the care team (level 2); organization (level 3); and environment (level 4) of the IOM Model of Care

Unlike the healthcare industry, quality improvement and quality engineering principles have long been established in the manufacturing and service industries. The foundation for total quality management (TQM) and continuous quality improvement (CQI) was laid out in the mid-1920s by Walter Shewhart, W. Edwards Deming, and Joseph Juran (Best & Neuhauser, 2006). Walter Shewhart, also known as the father of statistical control, advocated for management process to streamline production processes. He created the basis for control charts (Shewhart, 1931) and developed the Shewhart learning and improvement cycle which consists of: Plan, Do, Study, and Act (PDSA) (Best & Neuhauser, 2006). Like Shewhart, W. Edwards Deming during the same era made profound contributions to the quality movement. He developed a theory of management

based on 14 points of management that transformed style of management in industry, education and government towards optimization (Deming, 1986). Championing his work and that from his long collaboration with Shewhart, Deming later helped transform the Japanese manufacturing industry into one of the most competitive industries in the world, producing high-quality and innovative products (Aguayo, 1991). Joseph M. Juran is also an important contributor to the quality movement. Independent from Deming, Juran contributed to the success of the Japanese economy through his work in total quality management (TQM) and quality engineering (Godfrey, 1999).

Avedis Donabedian, a physican, is considered to be the founder of the study of quality in healthcare and medical outcomes research (M Best & D Neuhauser, 2004; Avedis Donabedian, 1988b; Mullan, 2001). In 1966, he developed the Donabedian model, a conceptual model that divides quality of healthcare measures into structure, process, and outcome (Avedis Donabedian, 1966; A Donabedian, 1980; Zimmerman, 2003). His model was used in diverse healthcare settings to modify structure and process with the objective of improving care services and health outcomes. Other quality improvement methodologies that were widely used and successful in the manufacturing and service industries did not gain traction in the healthcare industry until 1989 when CQI concepts and principles were introduced in healthcare (Berwick, 1989a; Laffel & Blumenthal, 1989). By 1993, based on the national survey of quality improvement activities in hospitals, 69% of the 3,300 hospitals responding to the national survey had adopted quality CQI and TQM programs (Barness et al., 1993). In 2012, a systematic review of the effectiveness of quality improvement (QI) methodologies adopted from manufacturing for surgical care indicated that QI methodologies can improve surgical

care in a wide range of areas including operating room efficiency and infection rate reduction (Nicolay et al., 2012). The systematic review included studies that used major QI methodologies adopted from manufacturing: CQI; Six Sigma; TQM; statistical process control (SPC) or statistical quality control (SQC); PDSA cycle; Lean; and Lean Six Sigma. Only one reported study tested the impact of QI interventions with a cluster randomized design; the others used non-randomized designs. The authors called for more rigorous randomized multicenter studies to help establish evidence-based management in healthcare.

Samsa and Matchar addressed whether, when, and how CQI can be assessed using randomized trials (Samsa & Matchar, 2000). The authors concluded that randomized trials are appropriate for studying CQI as conceptual models for generating intervention when the process improvements are costly, when care outcomes are substantial, and when the best approach to achieve the improvement cannot be easily established. Furthermore, the authors encouraged researchers to give consideration to the unit of randomization and whether the focus is on a specific intervention or on CQI as a conceptual model for generating interventions in healthcare. For example, organizational change interventions can be generated from conceptual models from behavioral sciences such as the transtheoretical model which posits that behavioral change goes through pre-contemplation, contemplation, preparation, action and maintenance stages (Prochaska, Prochaska, & Levesque, 2001); and decisional balance models which posit that movement of behavior across states is mediated by perceived benefits of behavioral change (Janis, 1977). In the cluster randomized trial used as a case study in this dissertation (discussed in detail in Chapter 3), several

behavioral change frameworks including Bandura's social cognitive theory (Bandura & Kazdin, 2000) were synthesized to study an intervention given to faculty in academic medicine, engineering, and science departments to promote gender equity behaviors (Carnes et al., 2014).

In general, CRTs in health systems engineering and health services research have been used to evaluate a wide range of healthcare quality improvement interventions (Garrison & Mangione-Smith, 2013) such as: implementation of electronic systems for drug reconciliation with medical team as the unit of randomization (Schnipper, Hamann, Ndumele, & et al., 2009); use of patient-specific electronic receipt of recommended care for patients with primary care team as the unit of randomization (Sequist et al., 2005); operationalization of initial care protocols in emergency departments with hospital as the unit of randomization (Yealy et al., 2004); use of improved collaboration practice between physicians and pharmacists for patient care with clinic as the unit of randomization (Carter et al., 2008); and community interventions intended to improve population health outcomes (Epstein et al., 2011; Group, 1995a, 1995b; O'Brien et al., 2007; West, Katz, Khatry, LeClerq, & et al., 1999). Furthermore, CRTs can be used to evaluate interventions that promote institutional or organizational change through workforce development (Carnes, et al., 2014; Grimshaw, Eccles, Marion, & Elbourne, 2005; Sequist et al., 2010).

Currently, one of the largest studies of organizational change and quality improvement ever conducted in health care was a cluster randomized trial that included 201 out-patient addiction treatment clinics in 5 states (Gustafson, et al., 2013; Quanbeck et al., 2011). The objective of the study was to find the most effective elements of health

collaborative in reducing wait times, rates of admissions and continuation in treatment as primary outcomes. The secondary outcomes included: treatment completion rates, level of adaptation and sustainability of recommended practice, organizational readiness to adopt and sustain the new practices, voluntary employee turnover, and program margin. Both the primary and secondary outcomes represent quality improvement and organizational change elements similar to those that have been traditionally addressed by industrial engineers in the manufacturing domain. This study confirms the value of using cluster randomized studies to rigorously evaluate the impact of healthcare quality improvement initiatives within health systems. Therefore, to become full partners and leaders in healthcare quality improvement, it is vital that health systems engineers become familiar with the intricacies of design and data analysis of cluster randomized trials.

1.3. Research Context

One of the design objectives of randomized trials is to generate comparable experimental groups by achieving a balanced distribution of baseline covariates. Covariate imbalance may be a result of chance, of selection bias, or of individuals not providing data at baseline (non-participation). Restricted randomization, such as stratification and pair matching, has been used in some situations as a design strategy to minimize the risk of baseline covariate imbalance. In spite of these strategies, the core objective of randomization may not be achieved when randomizing at a cluster level because baseline covariates of individuals within a cluster may be imbalanced (Leyrat, et al., 2013).

CRTs that depend on longitudinal survey response remain especially at risk of bias due to differential response rates across clusters over time. Responses may have a monotone missing pattern, in which subjects are censored at any given point in time, or may have a non-monotone missing pattern, in which subject responses are missing intermittently over time. Having differential attrition of responses across groups is not unusual for CRTs that depend on longitudinal survey responses. In a study in which 36 general practice clinics were enrolled to assess a redesigned community postnatal intervention, 4.2% (46/1087) in the intervention group compared to 2.6% (25/977) in the control group withdrew from or dropped out of the CRT (MacArthur et al., 2002). The differential attrition was found to be statistically significant $(p = 0.04)$. Similarly, differential attrition occurred in a study of the effect of Deltamethrin-impregnated dog collars worn by domestic dogs on protecting children from infection in which 11.4% (229/2006) and 7.6% (143/1870) dropped out of the intervention and the control groups respectively $(p < 0.001)$ (Gavgani, Hodjati, Mohite, & Davies, 2002).

Propensity scores, in such cases, can be introduced into the analysis of CRTs as a type of adjustment method (Giraudeau & Ravaud, 2009). Propensity scores, first published by Rosenbaum and Rubin, are defined as the conditional probability of an assignment to the intervention group given a vector of observed baseline characteristics (Paul R. Rosenbaum & Rubin, 1983). Propensity scores are commonly used to allow observational studies to inherit the characteristic of randomized trials (D'Agostino, 1998; Williamson, Morley, Lucas, & Carpenter, 2012). Currently, such observational studies may include the utilization of big data. However, it was not until 2013 that the use of propensity scores during the analytical phase of CRT with selection bias was

investigated (Leyrat, et al., 2013). In this simulation study, six analytical methods for estimating treatment effects were compared for bias and standard error. The six methods included an analysis with the following: propensity scores adjusted for as model covariates; data stratified on propensity scores; pair-matched design using propensity scores; weights defined as the inverse of propensity scores; covariate adjustment in regression; and no adjustment to a regular regression model. The investigation by *Leyrat et al.* did not include the impact of bias on differential attrition, or consideration of information inherent in subsequent missing data patterns over time, a likely scenario in CRTs with a longitudinal component. Furthermore, weights were applied to only individual-level data, omitting cluster level weights. This dissertation addresses this previously unexamined issue.

1.4. Research Objective

The objective of my research is to compare bias and efficiency between different methods of adjusting for covariate imbalance at baseline and attrition between cluster randomized groups analyzed by mixed effects linear models. This will be accomplished by clearly formulating the options in the setting of cluster randomization with longitudinal follow-up, applying the methods to a sample dataset, and simulating clustered longitudinal data with differential attrition.

The specific aims of the work of this dissertation are:

1. To develop several approaches to computing propensity scores and/or performing covariate adjustment for clustered longitudinal mixed models: by incorporating variables to account for both differential participation leading to the imbalance in

measured and unmeasured covariates between treatment groups, and by incorporating both measured covariates and indicators of patterns that adjust for dependence of longitudinal attrition on random effects.

2. To develop approaches to most appropriately incorporate propensity scores into the analysis.
3. To apply the above methods to a case study.
4. To compare, via simulation studies, the ability of the methods to reduce bias and the standard errors of estimators in settings generated to mimic real life examples.

Although mixed models are a natural way to analyze CRTs, previous research is limited to two-level models. Many studies depend on several follow-up points and therefore involve more than two levels. Also, propensity scores will be compared in several ways for longitudinal data in the context of CRTs. This study has the potential to make practical and vital contributions to the body of analytical methods that address specific types of biases which may affect the quality of statistical inference in CRTs. A wide range of industrial engineering methodologies are essential for understanding the complex interactions of healthcare subsystems, implementing healthcare systems quality improvement, and ultimately transforming healthcare. In addition to other intervention or system evaluation methods, it is important for health systems engineers to master design and analytic strategies for the most appropriate use of CRTs in healthcare redesign for the 21st Century.

CHAPTER 2
LITERATURE REVIEW

The primary goal of the chapter is to highlight literature that provides context to the historical development of quality improvement movement and the emergence of cluster randomized trials and the methodological challenges associated with this experimental design. This chapter is organized into five sections: synopsis of quality improvement in healthcare, junctures of statistical development in healthcare research, current methodological challenges, mitigation of selection and differential attrition bias and research focus.

2.1 Synopsis of Quality Improvement in Healthcare

A long history of efforts to improve the quality of care has been chronicled. One such notable effort can be traced to Ignaz Semmelweis, known as the father of infection control (Chassin & Loeb, 2011). Semmelweis, in 1847 during his appointment in a teaching hospital in Vienna, proposed the practice of washing hands with chlorinated lime solutions after he observed that the post-delivery mortality rate of women under the care of physicians and medical students was 13-18% compared to that of women under the care of midwives (2%) (Mark Best & Duncan Neuhauser, 2004). He believed there was an association between performing autopsies, as physicians and medical students did before attending to the women patients, and high mortality rate.

Florence Nightingale, a pioneer in evidence-based nursing, through her collection of works, epidemiological approach, and use of statistics made important contributions towards the establishment of a public healthcare system Britain (Chassin & Loeb, 2011; McDonald, 2001; Neuhauser, 2003; Palmer, 1977; Spiegelhalter, 1999). Nightingale established her plans for military medical education based on her experience in and analysis of the army medical units in Crimea in the 1858 publication *Notes on Matters affecting the Health, Efficiency and Hospital Administration of the British Army* (Attewell, 1998).

The roots of outcomes management and quality assurance in patient care can be traced to Ernest Codman (Avedis Donabedian, 1988a). American College of Surgeons (ACS), which was founded in 1913, adopted Codman's "end results system of hospital standardization" a quality assurance system in which patient outcomes are tracked post-treatment as a quality assurance system to improve quality of care (Maxwell, 1984; McIntyre, Rogers, & Heier, 2001; Shackford, 2006). The standardization system set a minimum standard of care.

In 1951, the American College of Physicians, American Hospital Association, American Medical Association, and Canadian Medical Association joined ACS in creating the Joint Commission of Hospitals (JCHA) (Roberts, Coale, & Redman, 1987). JCHA was charged with the providing voluntary accreditation for hospitals. In 1966 JCHA, which was renamed to Joint Commission on Accreditation of Healthcare Organization (JCAHO) in 1987, abandoned the minimum standard of care model in favor of the Donabedian's quality of care framework which has three dimensions of care: structure, process and outcome (Avedis Donabedian, 1966; Luce, Bindman, & Lee, 1994) . The

advent of Medicare in 1965 especially accelerated development and implementation of a myriad of quality initiatives, albeit from different perspectives, by other organization besides JACHO including government regulatory programs, utilization review committees, experimental medical care organizations, professional standards and review organizations, peer review organizations, and legislation bodies among other organizations (Chassin & Loeb, 2011; Luce, et al., 1994).

The contemporary quality improvement movement in healthcare has mostly benefited from ideas developed in health services research and modern quality improvement (T. A. Brennan, 2002). Health services research as field, which is considered to have been established between the 1950s and 1960s, is informed by a long history of cumulative knowledge on quality, access, delivery, utilization, outcomes and cost of care. In a 1952 seminal conference on research requirements for healthcare, J. N. Morris, from the UK Medical Research Council, presented broad applications of epidemiological concepts to medical services that extensively influenced emergence of health services research as a new field (McCarthy & White, 2000; White, 2002). Subsequently, research in health services was funded by a diverse set of organizations.

The 1946 Hill-Burton Act or Hospital Survey Construction Act, a federal law that was designed to fund the building or modernization of hospitals in underserved communities across the country, was expanded in 1955 providing the first funding mechanism for health services research (Bice, 1980; McCarthy & White, 2000). In 1966, the establishment of a federal government health services research study section led to the recognition of the use of term "health services research". In 1969, the field was allocated more funding after the establishment of National Center for Health Services

Research and Development (Kathleen N Lohr & Steinwachs, 2002). Through the 1970's funding for health services research continued in public and private organization domains. Federal funding through agencies including the National Center for Health Services Research (NCHSR) was used to support health services research (T. A. Brennan, 2002). The Rand Health Insurance Experiment (HIE), one of the most influential and largest health care financing experiment that was carried out from 1971 to 1982, was designed to answer how much more medical care people would use if it was provided free of charge and the consequences of utilizing the care on their health (Kathleen N Lohr et al., 1986). Health services research continues to be funded through grants and fellowships from major government agencies, from philanthropists, and from profit and non-profit organizations. Other significant studies such as the 1984 Harvard Medical Practice Study (MPS) demonstrated the need for implementation of quality improvement in the healthcare. In the MPS study, 30,121 randomly selected patient records from hospitals in New York State were examined for incidence of adverse events and negligence (T. A. Brennan et al., 1991; Leape et al., 1991).

Donald Berwick is largely credited for combining theories from health services research and modern quality improvement concepts and introducing them into medical care (Berwick, 1989b, 1991; T. A. Brennan, 2002). The evolution of modern QI methods was arguably triggered in the mid-1920, the foundation of which can be traced to Walter Shewhart, a physicist, engineer and statistician; W. Edward Deming, a statistician; Joseph Juran, a management consultant and engineer; Feigenbaum, an engineer and quality control expert, among others (T. A. Brennan, 2002; Flynn, Schroeder, & Sakakibara, 1994; Yong & Wilkinson, 2001). Berwick and Godfrey, in 1987, were co-

investigators of the National Demonstration Project (NDP) in Quality Improvement in Heath Care, in which 21 health care organizations, with the support of leading quality organizations in industry, enrolled to investigate the applicability of quality improvement methods in manufacturing in the health care environment (Berwick, Godfrey, Roessner, Plsek, & Garvin, 1990).

The cumulative effect of research and demonstration projects on the need for quality in healthcare influenced the publications of IOM reports *To Err is Human: Building a Safer Health System* (Kohn, et al., 2000; Leape & Berwick, 2005) and *Crossing the Quality Chasm: A New Health System for the 21st Century* (Berwick, 2002). The publication of these reports has fostered collaboration between IOM and National Academy of Engineers to transform health care (Reid, et al., 2005; Varkey, Karlapudi, & Bennet, 2008). The impact of modern quality improvement methods in health care has been largely successful. However, the health care environment is acknowledged to be more complex than that in manufacturing (De Souza, 2009). For example, patients may be assisted by a team or care providers with different specialization across different care settings such as acute care and long term care settings. The UK Medical Research Council has framework and guidelines for the design and evaluation of health care interventions (Anderson, 2008; M. Campbell, et al., 2000). Furthermore, CRTs are considered to be one of the suitable quantitative evaluative designs for complex interventions in health care (Craig, et al., 2008a; Loeb, 2002). The increasing recognition of the role of CRTs in evaluating quality improvement interventions emerging in the complex healthcare environment highlights the need for continued development of analytical methodology of this experimental design.

2.2 Junctures of Statistical Development in Healthcare Research

2.2.1 Comparable Groups, Randomization, and Experimental Methods

Controlled trials that randomize individuals have a relatively longer history than CRTs. The first written account of a 10 day comparative study dates back to 600 B.C. when Daniel of Judah compared the health effects of a vegetarian diet to the health effects of a royal Babylonian diet (Greenfield, 2004; Grimes, 1995). However, it was not until 1364 A.D. that the statistical concept of comparable groups in studies was recorded in a letter written by Francisco Petrarch to a fellow poet (Chalmers, Dukan, Podolsky, & Smith, 2012). In his letter, Petrarch affirmed his belief in the importance of comparable groups in establishing the effects of medicine. However, it was not until several centuries later that James Lind, a Scottish physician, conducted what has been widely accepted as the first experiment using concurrent treatment and control groups (Chalmers, et al., 2012; Kaur, 2013). Prior to embarking on the experiment that established that citrus fruits prevented and cured scurvy, he noted that the six groups required for the study were comparable. His paper, A Treatise on Scurvy, was published in 1753 (Lind, 1757).

Van Helmont in 1648 and Starkey in 1658 are recorded to be among the first to propose the concept of random assignment with the objective of removing bias; they randomly assigned patients to one of two physicians for intervention using a mechanism known as “casting lots” with the objective of minimizing bias (Chalmers, et al., 2012). Ronald Aylmer Fisher, in 1925, pointed out the importance of randomization in experimental design in his book Statistical Method for Research Workers. (Hall, 2007; Preece, 1990). Fisher’s 1935 publication, *The Design of Experiments* (DOE),

modernized agricultural research and later industrial, biological and medical research (G. E. Box, 1989; Ronald Aylmer Fisher & Yates, 1949; Thompson, 1990; Yates, 1964). DOE methods were successfully applied in the cotton Industry in the 1930's (G. Box, Bisgaard, & Fung, 1988) and in the military and other industries in the 1940s (Davim, 2012). Edward Deming, in the 1950's, introduced design of experiments to Japanese engineers and scientists (John & John, 1971). Geichi Taguchi beginning in the 1950's also developed quality improvement methodologies in Japan, including fractional factorial designs and other methods considered to be controversial (G. Box, et al., 1988). From the 1940's onward DOE methodologies for quality improvement permeated industry in the United States and Japan.

Randomization was not used specifically in clinical research until 1946 when the UK Medical Research Council carried out the MRC streptomycin trial of treatment for tuberculosis (Baron, 2012; Yoshioka, 1998). Austin Bradford Hill, the statistician on the MRC streptomycin trial, is widely credited for bringing an experimental approach into clinical medicine (Kunz, Vist, & Oxman, 2007). Prior to the streptomycin trial, the MRC investigated the efficacy of patulin treatment for the common cold between 1943 and 1944 in a double-blind quasi-randomized control trial. However, the streptomycin trial carried out between 1946 and 1947 is widely accepted as the first double-blind placebo-controlled randomized clinical trial.

2.2.2 Randomized Controlled Studies

Methodological advancement in and application of randomized controlled studies has grown rapidly since the 1950s. The widespread use of randomized controlled studies in medical research between 1950 and 2007 was illustrated by Bastian et al. who, using a

variety of data sources including the Cochrane Central Register of Controlled Trials (CENTRAL) and the Medical Literature Analysis and Retrieval System Online (MEDLINE), found that the number of trials published per year rose from less than 1000 per year to over 24,000 per year or on average approximately 75 trials per day (Bastian, Glasziou, & Chalmers, 2010). Randomized controlled studies are currently considered the gold standard; however, this view has been critically examined due to the influence of unpredictable human consciousness and behavior on bias (Kaptchuk, 2001). While there has been rapid methodological advancement for RCTs, methodological development of CRTs has continued to lag behind.

The split-plot design, invented by R. A. Fisher in 1925, is the earliest known experimental design in which whole blocks (clusters) are randomized into treatment and control groups (Jones & Nachtsheim, 2009). While the split-plot design has its origins in the agricultural experiments, it has been extensively used for industrial experiments that are characterized by factors that can and that cannot be manipulated with ease (G. E. Box, Hunter, & Hunter, 2005). The earliest examination of what we now refer to as CRTs was arguably initiated in 1940 by Everet F. Lindquist (N Klar & Donner, 2004). Lindquist advocated the use of clusters in educational research of students within classrooms. He emphasized the need to take into account clustering effects (correlated data within clusters) in the analysis. The development of rigorous statistical methods by Hansen and Hurwitz and then by Walsh considering the effects of correlated data within clusters emerged in 1942 for binary outcome data and in 1947 for Gaussian outcome data respectively (Cornfield, 1978; N Klar & Donner, 2004). In 1952, the imprecision of CRTs was addressed by Mainland in medical literature (N Klar & Donner, 2004). Even

though there were notable CRTs conducted in healthcare research, it was not until 1978 that an article by Cornfield on the limitations of CRTs reached a wider audience in health sciences (A. Donner & Klar, 2002; Murray, Varnell, & Blitstein, 2004). Cornfield's publication marked the beginning of an era of methodological developments in the design and analysis of CRTs. In1998 the first text book on CRTs was published (Murray, 1998).

The results of the first part of a bibliometric survey that reviewed the number of CRTs published in 1983, 1988, 1993, 1998 and 2003 showed less than 3 studies (methods or trials) in 1990 (Bland, 2004). In 2003, Bland found over 60 trials and over 20 methods articles. The results also indicated increased awareness of the impact of clustering on statistical analyses of the resultant data. The second part of the bibliometric survey consisted of a hand search of trials published in the British Medical Journal (BMJ) in selected years over a span of 20 years. The survey results confirmed an increase in the number and in the quality of CRT studies.

The methodological development and appropriate application of CRTs advanced slowly compared to individually randomized trials. A study of 16 non-therapeutic cluster randomized trials published between 1979 and 1989 examined the occurrence of six methodological issues associated with cluster designs. The authors found that less than 1/5 and less than 1/2 incorporated clustering in the sample size calculation and analysis, respectively (Allan Donner, Brown, & Brasher, 1990). A methodological review of primary prevention trials published between 1990 and 1993 in the American Journal of Public and Health and Preventive Medicine found that out of 21 trials, 19% and 57% allowed for clustering in the sample size calculation and in the analysis,

respectively (Simpson, Klar, & Donnor, 1995). In a systematic review of 152 trials in primary care between 1997 and 2000 indexed in the Cochrane Central Register of Controlled Trials (CENTRAL) and in the UK National Register as well as unpublished trials submitted in conference proceedings, 9% accounted for clustering in establishing a sample size and 59% accounted for clustering in the analysis (S. M. Eldridge, Ashby, Feder, Rudnicka, & Ukoumunne, 2004). The unpublished trials were more recent and adhered to a higher standard of quality.

2.3 Current CRT Methodological Challenges

CRTs are regarded as an important tool in the evaluation of interventions in health services research (M. J. Campbell, Donner, & Klar, 2007), and considered to be one of the experimental designs suitable for evaluating complex interventions (Craig et al., 2008b). The definition of complex interventions has broadened and moved beyond being characterized by multiple components (Anderson, 2008) to include: the number of interacting components, the number and difficulty of behavioral changes needed for a successful intervention, the number of organizational levels targeted for the intervention, the number and variability of outcome variables, and the degree to which an intervention can be tailored (Craig, et al., 2008b). The characteristics of quality improvement interventions, implemented in the context of health systems engineering, often align with the definition of complex interventions.

The CRT is an appropriate tool for evaluating interventions that cannot be directed toward selected individuals but are instead delivered to aggregates of individuals who are defined as a cluster or subsystem (A. Donner & Klar, 2002; Fayers, Jordhøy, &

Kaasa, 2002; Murray, 1998). This creates a nested hierarchical randomized controlled experimental design in which individuals are nested within clusters, which are in turn nested within the treatment groups. Although CRTs have some clear benefits and are indeed necessary when studying an intervention in a complex system, Walsh in 1947 and Cornfield in 1948 identified the statistical inefficiency of randomization at the cluster level which results in fewer replications of the treatment than in individual randomized trials (Cornfield, 1978; Walsh, 1947).

Members of a cluster are more likely to have similar outcomes than would a random sample from a population, especially in the case when outcomes are related to the characteristics of a cluster (S. Eldridge & Kerry, 2012; Lewsey, 2004). The lack of independence of individuals in the same cluster creates within-cluster homogeneity. The underlying differences between clusters vary by trial. In some trials, for example, populations that serve as units of randomization may have similar socio-economic status while in other trials populations may share demographic characteristics that relate to responses. Features of the intervention may make possible the sharing of experiences within clusters creating clustering effects (Koepsell, 1998).

Empirical investigations are usually required to establish the reasons for between-cluster variability. The intraclass correlation (ICC) quantifies the proportion of between-cluster variance out of the total variance. Statistical efficiency is a function of the variance inflation factor (design effects) resulting from clustering, which in turn is a function of the ICC and the average cluster size (Hsieh, Lavori, Cohen, & Feussner, 2003). Since the effective sample size is less than the total number of individuals, statistical power can be estimated by multiplying the sample size by the variance

inflation factor $1+(m-1)\rho$ where m and ρ are defined as the average clusters size and ICC respectively (A. Donner & Klar, 2002; Murray, et al., 2004). The within cluster dependencies need to be accommodated both in the design and analysis of the CRT. Considering the loss of statistical efficiency, implementing a cluster randomized design should be motivated by compelling factors such as the nature of the intervention being a cluster level process, infeasibility of delivering the intervention at an individual level, the need to minimize the threat of contamination bias, ethical considerations, administrative costs, and the value of the ICC among other considerations (M. K. Campbell, et al., 2004).

There is a tradeoff between the loss of precision in cluster randomization and the increased risk of contamination bias in individual randomization in context of the sample size and subsequently the cost and complexity of the experiment. (Slymen & Hovell, 1997; Torgerson, 2001). Minimizing the risk of contamination bias across individuals in different treatment groups enhances participant compliance (N. C. Campbell, et al., 2007; A. Donner & Klar, 2002). For example, if the intervention is at the provider level and the outcomes are measured both at the patient and provider level, then contamination is significantly reduced when randomization is at the physician (cluster) level. In this case, the physicians and their panels are assigned to the same experimental group. On the other hand, if the individual patients are randomized, then physicians, who are receiving the intervention, may treat both intervention and control patients.

In CRTs the unit of randomization and unit of inference need not coincide. Because the implications of clustering in the design and analysis of cluster trials were sometimes ignored, reporting guidelines specifically for CRTs were published as an extension of the Consolidated Standard of Reporting Trials (CONSORT) (M. K. Campbell, et al., 2004) that included required flow diagrams. In 2011, Ivers et al. conducted a review comparing the methodological and reporting qualities of 300 randomly sampled cluster randomized trials published before 2000-2004 and after 2005-2008 when the 2004 CONSORT guideline was extended to CRTs (N. Ivers et al., 2011). The authors found that, while there was some improvement in the quality of reporting CRTs, adherence to methodological guidelines continued to be inadequate. If the unit of inference is the individual, statistical analyses that do not incorporate clustering increase the likelihood of committing a type I error which occurs when the null hypothesis is erroneously rejected (i.e., results infer a significant treatment when none exists), hence inflating the statistical significance of the results (S. Eldridge & Kerry, 2012; Zucker, 1990). The lack of compliance may be attributed to a lag in the methodological development and in the dissemination of guidelines associated with CRTs as an experimental design.

When the unit of randomization and the unit of inference are both at the cluster level, there is no need to adjust for clusters in the analysis. In this case, the CRT becomes equivalent to the standard clinical trial. For example, the intervention in a CRT testing the effect of guidelines on the reduction of inappropriate referrals for radiographic examination was directed at the practice (cluster) level (Oakeshott, Kerry, & Williams, 1994). The total of 62 practices (unit of allocation) that were randomized into intervention and control were the unit of inference rather than the individual patient. In

most cases, the unit randomization is cluster and unit of analysis the individual within a cluster. Consequently, clusters must be accounted for in the design and analysis of the study.

Assuming adherence to CONSORT guidelines, randomization of clusters, similar to randomization of individuals can create comparability of the experimental groups. It is important to balance baseline covariates between the groups (McEntegart, 2003; Pocock & Simon, 1975). Comparability of groups is achieved if the outcome variable, when conditioned on the allocation, is independent of the group (Abel & Koch, 1999) (i.e., the same treatment effects would have been observed reversing the randomization); as pointed out by Abel and Koch, randomization does not imply comparability but rather distributional equality of known and measurable variables. Hence, comparability, from this perspective, is a theoretical construct. Cluster randomization as result of ICC should be given careful consideration to avoid biased allocation (A. Donner & Klar, 2002). The imbalance of baseline covariates across intervention groups decreases statistical power and precision of the trial bringing into question the face validity credibility of the trial results (Perry, Faes, Reelick, Olde Rikkert, & Borm, 2010; Pocock & Simon, 1975; Therneau, 1993). CRT baseline covariates are particularly at risk of imbalance in trials with a few clusters because individuals embedded in the clusters are not randomized. For example, baseline distributions are imbalanced in a study that enrolls only 8 clinics (clusters) if all 3 of the 8 clinics that serve predominantly low-income populations are randomized to the intervention group along with 1 clinic that serves an affluent population. A review of CRT trials conducted between 2000 and 2008 found that while the median number of

clusters was 21, 25% of the studies had less than 12 clusters and 14% of the studies had less than 4 clusters per intervention group, the fewest recommended (N. Ivers, et al., 2011; N. M. Ivers et al., 2012).

The risk of selection bias resulting from the enrollment of a small number of clusters or from differential recruiting may undermine internal validity of the study. The Cochrane handbook for systematic reviews of interventions defines internal validity as the extent to which a study prevents systematic bias through good design, conduct and analysis of the study. (Higgins & Green, 2008). Selection bias generates imbalances in the distribution of covariates between experimental (or experimental and control) groups. In some instances, identifying the subjects prior to randomization is not feasible, leaving open the possibility of differential recruitment into the clusters to be studied. For example, after the randomization of clinics, if physicians are responsible for selecting patients for treatment, then physicians in the intervention group may enthusiastically select to treat patients they think are likely to benefit most from the intervention. In a systematic review of cluster randomized trials published between 2004 and 2005, 25% of the trials showed evidence of susceptibility to selection bias resulting from differential recruitment (S. Eldridge, et al., 2008). In a 2008 review of studies published in four leading medical journals, 5 out of the 24 cluster randomized trials demonstrated some recruitment bias (Brierley, Brabyn, Torgerson, & Watson, 2012) and in a separate study of 36 cluster randomized trials in three major general medicine journals, 7 out of 23 trials that had not identified subjects prior to randomization showed evidence of differential recruitment (Puffer, Torgerson, & Watson, 2003). As another source of selection bias, CRT trials with a longitudinal component are particularly vulnerable to

differential attrition. In the latter review of 36 trials, 4 had evidence of differential attrition (loss to follow-up) (Puffer, et al., 2003).

2.4 Mitigation of Selection and Differential Attrition Bias

From a design perspective, unrestricted cluster randomization in which clusters are simply randomized to intervention or control groups does not effectively minimize the risk of baseline covariate imbalance in CRTs as it does in individually randomized trials. This is because the number of randomized units (clusters) in CRTs is typically considerably smaller than that in individually randomized trials. Restricted randomization in which stratification or pair-matching of clusters prior to random allocation into intervention or control groups have been employed to mitigate the risk of covariate imbalance. In these designs, random allocation is carried out within each stratum or pair (M. K. Campbell, et al., 2004; N. Ivers, et al., 2011).

A stratum is defined by cluster-level baseline covariates that clusters have in common and that may influence distribution of baseline covariates. Clusters within each stratum are randomly allocated to the intervention and control groups, resulting in a replication of unrestricted randomized design within the stratum (A. Donner & Klar, 2002). The matched-pair design is a special case of a stratified design in which only 2 clusters (stratum) are matched by cluster-level baseline covariates and then randomly allocated to the intervention and control groups. Matched-pairs need to be comparable for matching to be effective (N. M. Ivers, et al., 2012). In the case of CRTs with a small number of matched-pairs, the loss of a single member of the pair necessitates dropping both clusters from the study and can lead to a substantial erosion of power (Diehr,

1995; Allan Donner, Taljaard, & Klar, 2007). Power is defined as the probability that the null hypothesis is rejected given that it is false. However, matched-pairs with a 0.45 correlation or higher between the pairs can have more power than unmatched studies as long as more than 5 pairs are used (Martin, Diehr, Perrin, & Koepsell, 1993). The potential for loss to follow-up in all CRT designs should be incorporated in the sample size calculations (Neil Klar & Donner, 2001).

There is growing interest in the benefits and limitations of restricted randomization in CRTs. Although CRTs require more participants than standard randomized trials to have the same power, randomization stratified by factors such as cluster size, selected demographic characteristics, geographical location, and health districts when used may increase the statistical power of a study (S. Eldridge & Kerry, 2012). To address the impact of restricted randomization on power in CRTs, Lewsy (2004) performed a simulation study comparing stratified randomization using cluster size as the stratification factor with completely unrestricted randomization. The simulation found that the stratified design yielded greater statistical power when cluster size is strongly associated with an important cluster level factor that is not included in the data analysis (Lewsey, 2004). In a review looking for evidence for risk of bias in CRTs published in 3 major journals between 1997 and 2002, the investigators found that 25 (69%) of the 36 trials identified employed stratified allocation (Puffer, et al., 2003). Between 2000 and 2008, 56% of CRTs utilized restricted randomization out of which 19% were pair-matched trials, 34% utilized stratification, and 4% used other restricted randomization methods (N. M. Ivers, et al., 2012). Despite use of restricted randomization, imbalances in the distribution of baseline covariates may persist due to an unmeasured or unknown

baseline covariate or due to chance (Puffer, et al., 2003). Other allocation techniques such as best balance allocation, in which clusters are divided into two groups in all possible ways and the optimal allocation is selected, have been developed to minimize imbalance of baseline covariates (De Hoop, Teerenstra, Van Gaal, Moerbeek, & Borm, 2012).

In spite of this attention to reduce bias from covariate imbalance at the design stage, CRTs that depend on the participation of individuals after the random allocation phase continue to remain at risk of bias due to differential response rates leading to differential attrition across groups (Gavgani, et al., 2002; MacArthur, et al., 2002). In an examination of large field trials (CRT), Feldman and McKinlay showed that a follow-up attrition of 5% per year, has an accumulated effect of reducing the cohort measured at baseline by approximately 25% within a 5 year period (Feldman & McKinlay, 1994). In some cases when attrition is unrelated to the intervention, oversampling may be a solution (A. Donner & Klar, 2002). Roy et al. developed a model for sample size determination for trials randomized at a cluster level with differential attrition rates (Roy, Bhaumik, Aryal, & Gibbons, 2007). However, this solution relies on the assumption that estimates of differential attrition trials are available at the experiment design stage.

2.5 Research Focus

This research focuses on the problem of imbalance at baseline as a result of some individuals not contributing data and of imbalance at follow-up as a result of differential attrition. The effects of covariate imbalance at baseline and of attrition on the outcome need to be modeled in the analysis of the resulting outcome. Propensity scores can be

used to remedy imbalance from baseline participation and attrition during follow up. Propensity scores were developed by Rosenbaum and Rubin as a means to balance covariates in the two groups (Paul R. Rosenbaum & Rubin, 1983). A propensity score is defined as the probability of being assigned to a given experimental group given a vector of observed baseline characteristics (Paul R. Rosenbaum & Rubin, 1983). Propensity scores have been proposed as summary covariate measures that can be used at the analysis stage by utilizing different adjustment techniques (Leyrat, et al., 2013). Propensity-score based methodologies have typically been used to design observational studies, similar to the way randomized experiments are designed, to balance distribution of covariates across treatment groups by applying techniques such as matching, stratification or weighting to minimize bias (d'Agostino, 1998; Dehejia & Wahba, 2002; Paul R Rosenbaum & Rubin, 1984; Donald B Rubin, 1997, 2001; Stukel et al., 2007). A propensity score for a given subject $i = 1,...,n$, defined as the conditional probability of being assigned to intervention $I_i = \{0,1\}$ where $0 = \text{control}$ and $1 = \text{intervention}$ given a vector X of observed covariates x_i, is denoted by:

$$s(x_i) = P(I_i = 1 \mid X_i = x_i) \cdot \tag{2.1}$$

Propensity score theory is based on three assumptions (Paul R. Rosenbaum & Rubin, 1983). In the first assumption, the intervention allocation and the covariates are assumed to be independent conditional on the covariates:

$$P(I_1,...,I_n \mid x_1,...,x_n) = \prod_{i=1}^{n} s(x_i)^{I_i} \left(1 - s(x_i)\right)^{1-I_i}. \tag{2.2}$$

The second assumption requires that all covariates that are related to both the intervention assignment I_i and r_i response are known. In the third assumption, the intervention assignment and the response are conditionally independent given the covariates. This is referred to as a strongly ignorable intervention assignment:

$$(r_0, r_1) \perp I \mid x \ , \ 0 < P(I = 1 \mid x) < 1 \quad \text{or} \quad (r_0, r_1) \perp I \mid s(x) \ . \tag{2.3}$$

The propensity score function is almost always unknown. However, a logistic regression model can be applied to observed data to estimate the probability of being in the intervention given covariates. This model is formulated as:

$$s(x_i) = \frac{e^{(X_i\beta)}}{1+e^{(X_i\beta)}} = \left[1+e^{-(X_i\beta)}\right]^{-1}, \quad \text{where} \quad X_i\beta = \beta_0 + \beta_1 X_{i1} + ... + \beta_{p-1} X_{i(p-1)}. \tag{2.4}$$

In this dissertation, I especially consider CRTs with longitudinal follow-up because propensity scores can be used to remedy both imbalances from baseline non-participation and from differential (data missingness over time), which further complicates the analysis. Mixed effects models are a natural choice for analyzing longitudinal observations or repeated measures. George Airy, a British astronomist, is credited for introducing the variance components model (Airy, 1861) while R. A. Fisher is credited with formalizing the theoretical framework in the context of Analysis of Variance (ANOVA) (Ronald A Fisher, 1919; S. R. A. Fisher et al., 1970). Their work initiated a long history of research that, by the 1980's, led to the emergence of modern mixed effects regression models that could be applied to data that have irregular measurement times and time-varying and time-invariant covariates (Fitzmaurice & Molenberghs, 2009) . The development of a unified approach to fitting a general family

of models was particularly important to the evolution of these modern mixed effects models (Laird & Ware, 1982; Molenberghs & Verbeke, 2005). Parameter estimation in such models is often achieved through maximum likelihood or restricted maximum likelihood.

The models will take the following general form:

$$y_{ijkt} = \beta_0 + \beta_1 \cdot I_{i=1} + \beta_2 \cdot I_{t=t_2} + \beta_3 \cdot I_{t=t_3} + \beta_4 x_{ijk} + \beta_5 \cdot I_{i=1,t=t_2} + \beta_6 \cdot I_{i=1,t=t_3} + \gamma_{ij} + \lambda_{1j} + \omega_{ijk} + \varepsilon_{ijkt}. \quad (2.5)$$

The outcome variable for the models is y_{ijkt} where $i = 0,1$, $j = 1,...,\mathrm{m}$, $k = 1,...,\mathrm{n}_j$ and $t = 1,2,3$ represent the t^{th} follow-up time for the k^{th} person in the j^{th} cluster in the i^{th} experimental group. The random effects (2.6) include: γ_{ij} variability between-clusters; λ_{1j} variability between the interaction cluster and experimental group to examine the variability in the treatment effect; ω_{ijk} variability between persons; and random error ε_{ijkt} is defined as σ^2:

$$\gamma_{ij} \sim N\left(0, \sigma^2_{b_1}\right), \quad \lambda_{1j} \sim N\left(0, \sigma^2_{b_2}\right), \quad \omega_{ijk} \sim N\left(0, \sigma^2_{b_3}\right), \quad \varepsilon_{ijkt} \sim N\left(0, \sigma^2\right). \quad (2.6)$$

Maximum likelihood inferences are valid under MAR, i.e. if an observation being missing depends only on observed values of outcomes and covariates included in the modeling. However, missingness may also depend on variables that are not in the model. Complication arises when data are neither MCAR nor MAR. The missing data mechanism in this case is referred to as non-ignorable, as likelihood inference ignoring missing values is no longer valid. In this case, conditioned on all available observed

information, missingness still depends on unobserved responses. Modeling assumptions can be specified in some situations to adjust even for such missingness for example by assuming that missingness depends on unobserved random intercepts or random slopes (Park, et al., 2002). For such situations, pattern mixture models have been shown useful and relatively easy to apply. As noted by Roderick Little (1995), such models specify conditional distributions of variables given their missingness status. One approach is to group individuals by the pattern of their missing observations (Hedeker & Gibbons, 1997; Little, 1993, 1995; Pauler, McCoy, & Moinpour, 2003). A novel idea in my research is to incorporate the pattern of missing data into propensity scores that will subsequently be applied to the analysis utilizing mixed effects models.

Combining the missingness pattern with missingness predicted by covariates into propensity scores for the CRT of an intervention delivered to academic departments (Chapter 3) will attempt to correct for both MAR and non-ignorable missingness. It is likely that data missingness in this trial is correlated with observed covariates such as faculty rank and gender which were incorporated into the models. Random-effects models using maximum likelihood estimation may then be validly applied as long as there are no other influences on participation and attrition (Hedeker & Gibbons, 1997; Laird, 1988; Laird & Ware, 1982). However, we have not yet explored whether missingness may also depend on random effects underlying the individual trajectories , which can be corrected for via the pattern approach (Park, et al., 2002).

This research will be the first to model propensity scores in a longitudinal CRT to assess their ability to mitigate selection bias and differential attrition. The proposed work will make an important contribution to health systems engineering because quality

improvement interventions in healthcare are increasingly being implemented and evaluated with CRTs that have longitudinal follow-up. The objective of this research aligns with the overarching objective of the Institute of Medicine and the National Academy of Engineering partnership which calls for the streamlining care delivery system and continuously improve quality (Chassin & Galvin, 1998; Reid, et al., 2005).

CHAPTER 3
RESEARCH METHODS

In this chapter, I will discuss different methods used to adjust for bias resulting from covariate imbalance at baseline and from attrition of participants over time. The method used to generate a cluster randomized trial dataset with longitudinal follow-up time will be discussed. The following topics will be subsequently considered: (1) the generation of datasets that simulate different types of attrition, (2) the different approaches used to estimate propensity scores, (3) the different approaches used to incorporate propensity scores into the analysis. The application of propensity scores will be carried out via simulation studies to evaluate the abilities of the methods to reduce bias. The R software for statistical computing (Team R Core, 2014) was used for this research. Specifically, the lme4 linear and mixed-effects model package (D. Bates, Maechler, & Bolker, 2012; D. M. Bates, 2010) and the lattice graphic package (Sarkar, 2008).

3.1 Research Objectives

The objective of this research is to test whether methods based on propensity scores can correct bias arising from baseline imbalance and/or differential attrition when applied to CRTs with longitudinal components. The specific aims of this research are:

1. To develop several approaches to compute propensity scores that will be subsequently included in data analysis using mixed effects models. This will be accomplished by incorporating variables that account for covariate imbalance and

for differential participation between intervention groups and by incorporating both measured covariates and indicators of patterns that adjust for dependence of longitudinal attrition on random effects.

2. To develop approaches to most appropriately incorporate propensity scores into the analysis as covariates and by stratification.
3. To apply the methods to a case study.
4. To use simulation studies to compare the effectiveness of the methods to reduce bias and to compare standard errors of estimators.

3.2 Research Methods

3.2.1 Dataset Construction

The data used to develop the methods for this research were generated using parameter estimates from a cluster randomized trial that was developed to test an intervention that breaks gender bias habits in academic medicine, science and engineering at the University of Wisconsin-Madison (Carnes, et al., 2014). The experiment, which is also used as a case study, is relevant to this research because it informs the choice of fixed and random parameters from which clustered data arising from a CRT can be generated. Furthermore, a university, like a health care system, has semi-autonomous multilevel units (departments) that are suitable for CRTs. Hence, this study provides the opportunity to test the analytical methodologies developed in this research.

3.2.1.1 Cluster Randomized Trial Case Study Description

The study was a pair-matched, single-blind, cluster randomized controlled trial that was undertaken between September 2010 and March 2012 at the University of Wisconsin-Madison. The objective of the study was to evaluate the impact of a gender bias habit-changing intervention. A total of 92 departments (clusters) participated. Descriptive statistics were used to pair departments based on the following criteria: school/college, division, size, percent of female faculty members in the department, percent of junior faculty in the department, and percent of faculty in the department considered to be clinical faculty as opposed to tenure track faculty. Of the matching variables, percent female and junior faculty could not be evenly matched in control and intervention groups; hence, they were included as control variables in the data analysis.

Prior to randomization, the study directors met with the departments to inform them about their inclusion into the study and that a workshop would be offered to them within a 3 year period. They were also informed that they would be invited to complete multiple online surveys and that multiple surveys were necessary to “calibrate” the survey instruments. The presentation by the directors to the department members can be assumed to be unbiased because the random allocation had not been initiated and because the departments were unaware that any randomization would take place. Subsequently, within each pair, departments were randomized to the intervention group or to the wait-list control group. Data were primarily collected from the study using an online survey instrument that was designed by the study directors. A random number generator was used to assign random numbers to all departments. The department in each pair that received the higher random number was assigned to the intervention group while the department with the lower number was assigned to the wait-list control

group. The allocation sequence, once assigned, was not changed. Hence, there were no deviations from the random assignments.

The intervention consisted of a 2.5 hour workshop to increase awareness of implicit bias and to transform participants from motivation to action via “impulse to change” (Figure 2), a process that requires both self-efficacy and positive outcome expectations (Bandura, 1977, 1991, 1997; Carnes et al., 2012).

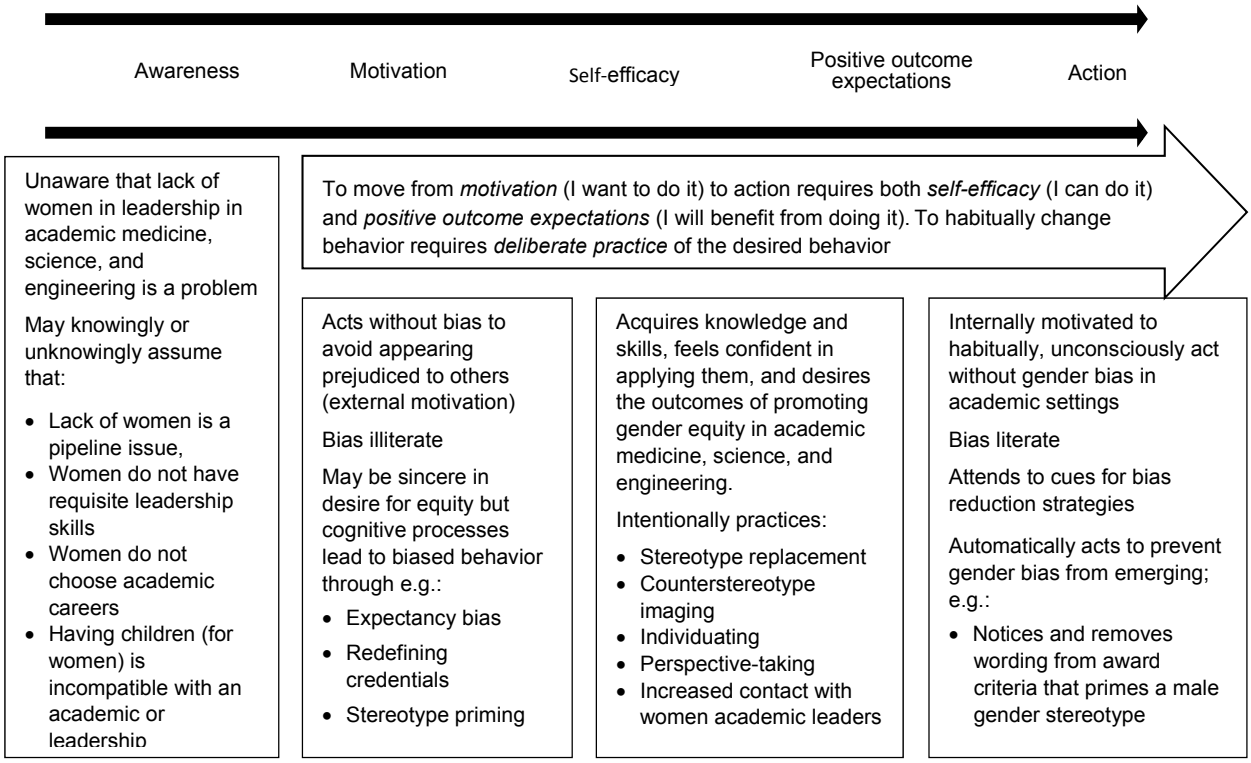

Figure 2: Individual Transformation from Awareness to Action (Carnes et al., 2013)

Measurement scales developed for the study fell into the following broad categories: Implicit Association Test (IAT); Awareness; Motivation; Gender Equity Self-Efficacy; Gender Equity Positive/Negative Outcome Expectations; and Action. The surveys were administered online by the Implicit Social Cognition Laboratory, at the University of Virginia. Only the self-efficacy scale will be used for this research to develop the

research dataset and to test the methods developed. Self-efficacy measures the extent to which individuals believe that they have requisite skills to enact gender equity (Bandura, 1977, 1991). Our self-efficacy measure consisted of five Likert scale items with seven levels ranging from strongly disagree (1) to strongly agree (7). The scale items were developed from a combination of themes drawn from focus groups with faculty and staff and from research in the areas of social cognitive theory and gender equity (Bandura, 1977, 1991; Eagly & Karau, 2002; Hill, Corbett, & St Rose, 2010; Isaac, Kaatz, Lee, & Carnes, 2012). The study design is provided in Figure 3.

Department faculty in both the control and intervention group were invited to fill out surveys at baseline and again at 3 days as well as 3 months after the workshop. While the unit of randomization was department (cluster), the unit of analysis was the individual. The treatment effect is measured as the mean difference minus the baseline difference between the intervention and control group at 3 day and at 3 month time points. This ensured that the existing baseline differences were accounted for in the evaluation of treatment effects.

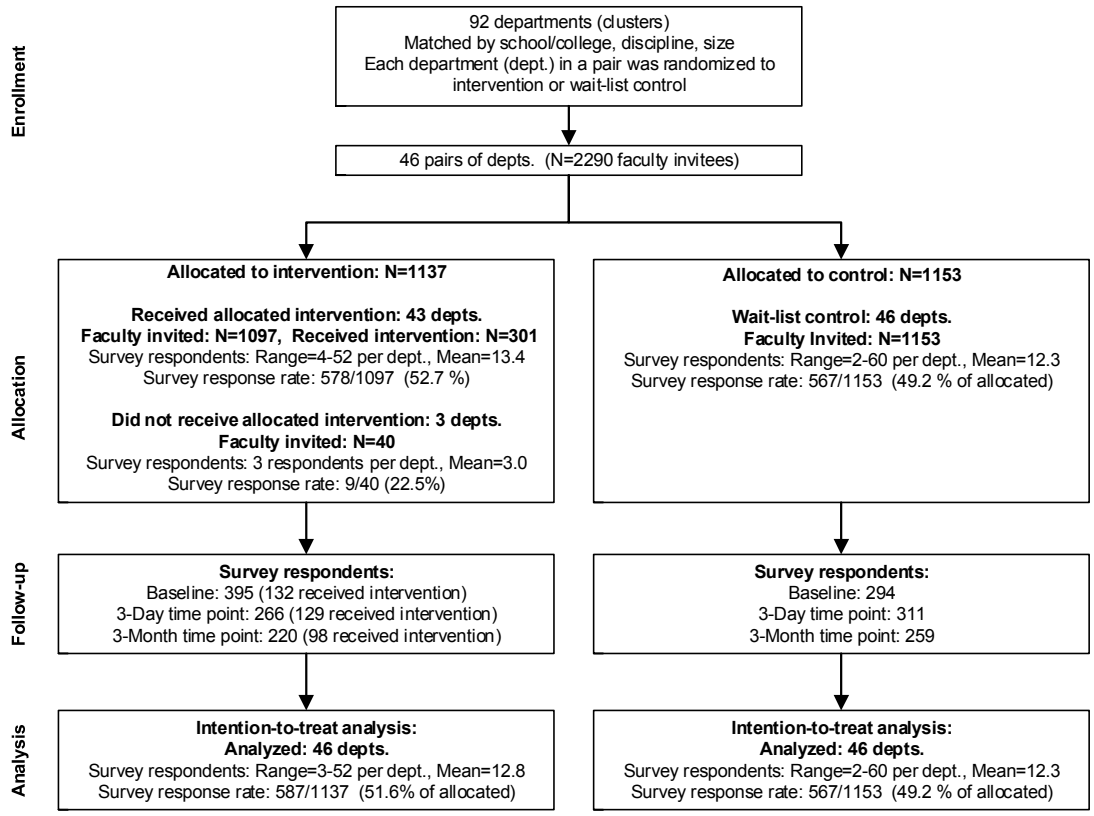

Figure 3: Flow diagram of gender bias-reduction cluster randomized trial presented as recommended by Consolidated Standards of Reporting Trials (CONSORT) statement extended to cluster randomized trials (M. K. Campbell, et al., 2004; Carnes, et al., 2014).

3.2.1.2 Dataset Construction Process

The bias literacy CRT study was used to provide data structure for simulating data by using model parameters from the trial. The data construction process was carried out in four stages. The first stage involved creating a dataset with no missing covariate information. The second stage consisted of the creation of records for any periods in which respondents did not submit a survey. The third stage applied simulation to generate efficacy outcome data using model parameter estimates from the bias literacy CRT efficacy scale. The fourth stage simulated datasets with different types of missing data mechanisms to test the bias adjustment methods developed in this dissertation.

Stage I

In the first stage of the data construction process, gender information not provided by some bias literacy study respondents was assigned to the individual. The process of assigning gender to an individual was a Bernoulli trial with the probability of being female (success) defined as the percent of female faculty members in the department to which the individual belongs. Self-reported gender in the bias literacy trial was as follows: 578 males, 378 females and 198 unreported. All cluster level information including the percent of females in the departments and percent of junior faculty in the departments was available for all respondents. Respondents are defined as faculty members/individuals who provided data in any of the survey periods (baseline, 3-Day follow-up, or 3-Month follow-up). To reduce the complexity of the investigation, the faculty rank status variable was not included in this study.

Stage II

During the second stage of the data construction, a new record was created for respondents that did not respond to all 3 surveys (baseline, 3-Day, and 3-Month). The objective of stage II was to ensure every respondent had all 3 records. The new record inherited all cluster level and individual level characteristics from existing responses. However, at this stage, the participant outcome/response for the new record was not filled with data; hence, the newly created record had outcome data that were missing.

Estimated models using the stage II dataset and bias literacy efficacy dataset were compared (Table 1). The parameter estimates were similar. The treatment effect is estimated by the interaction between the experimental group and the time period, which

is $\beta_{3-Day} = 0.17$ and $\beta_{3-Month} = 0.21$ for bias literacy data and $\beta_{3-Day} = 0.17$ and $\beta_{3-Month} = 0.20$ for the stage II data. Recall that the treatment effect is interpreted as the mean difference between the intervention and control group at follow-up (3-Day or 3-Month) minus the difference between the intervention and control group at baseline period.

Table 1: Parameter Estimates and Standard Errors from Bias Literacy and Stage II Data

	Bias Literacy Dataset	Stage II Data	Bias Literacy Dataset	Stage II Dataset
Fixed Effects	**Model Estimates**	**Model Estimates**	**SE**	**SE**
Intercept	5.39	5.39	0.06	0.05
Experimental Group	-0.04	-0.04	0.06	0.06
3-Day Survey Period	0.07	0.06	0.05	0.05
3-Month Survey Period	0.11	0.10	0.06	0.06
Gender (Female)	0.15	0.15	0.05	0.05
Interaction (Exp. Group and 3-Day)	**0.17**	**0.17**	**0.07**	**0.07**
Interaction (Exp. Group and 3-Month)	**0.21**	**0.20**	**0.08**	**0.08**

The random effects structure that will be used for this investigation differs slightly (Table 2 and 3). For this study, we ignore pairing as a clustering group in order to allow for generalization to CRTs with or without restricted randomization such as pairing, which is a more common scenario.

Table 2: Estimates from Bias Literacy Dataset

Random Effects - Groups	Variance
Person ID	0.4057
Interaction (Exp. Group & Pair)	0.0000
Pair	0.0132
Residual	0.2563

Table 3: Estimates from Stage II Dataset

Random Effects - Groups	Variance
Person ID	0.4183
Department	0.0000
Interaction (Dept. & Intervention)	0.0000
Residual	0.2564

Stage III

The objective of the third stage was to generate new survey response data for all dataset records using parameter estimates from the bias literacy trial. Prior to the generation of a complete dataset, the random effects parameters were first amplified to strengthen the clustering effects in the data. After the random effects were amplified, the following model was used to generate the complete dataset:

$$y_{ijkt} = \beta_0 + \beta_1 x_i + \beta_2 \cdot I_{t=t_2} + \beta_3 \cdot I_{t=t_3} + \beta_4 x_{ijk} + \beta_5 x_i \cdot I_{t=t_2} + \beta_6 x_i \cdot I_{t=t_3} + \gamma_{ij} + \lambda_{1j} + \omega_{ijk} + \varepsilon_{ijkt}. \quad (3.1)$$

The outcome generated is y_{ijkt}, where $i = 0,1$, $j = 1,...,m$, $k = 1,...,n_j$ and $t = 1,2,3$ represent the t^{th} follow-up time for the k^{th} person in the j^{th} cluster in the i^{th} experimental group. The bias literacy fixed effects parameter estimates (Table 1) and the amplified random effects (Table 4) were used to generate the outcome data. The fixed effects mean parameter estimates calculated from the 5,000 simulations (Table 5) are comparable to those of bias literacy CRT (Table 1) and the random effects mean variance parameter estimates calculated from the 5000 simulations (Table 6) are similar to amplified random effects used to generate the stage III data (Table 4).

Table 4: Amplified Clustering Parameters to Generate Outcome Data

Random Effects - Groups	Random Effects	Variance Used
Person ID	$\omega_k \sim N\left(0, \sigma^2_{b_3}\right)$	0.4000
Department	$\gamma_j \sim N\left(0, \sigma^2_{b_1}\right)$	0.4000
Interaction (Dept. & Intervention)	$\lambda_{ij} \sim N\left(0, \sigma^2_{b_2}\right)$	0.5000
Residual	$\varepsilon_{ijkt} \sim N\left(0, \sigma^2\right)$	0.2600

Table 5: Complete Dataset Parameter Estimate Means (5,000 Simulations)

Fixed Effects	Model Estimates	SE
Intercept	5.3803	0.6898
Experimental Group	-0.0348	0.7054
3-Day Survey Period	0.0625	0.0333
3-Month Survey Period	0.1031	0.0333
Gender (Female)	0.1460	0.0489
Interaction (Exp. Group and 3-Day)	0.1714	0.0464
Interaction (Exp. Group and 3-Month)	0.2047	0.0464

Table 6: Complete Dataset Variance Estimate Means (5,000 Simulations)

Random Effects - Groups	Variance
Person ID	0.4000
Department	0.4020
Interaction (Dept. & Intervention)	0.4933
Residual	0.2563

Stage IV

The objective of the fourth stage was to simulate different missing data mechanisms. Data missing completely at random (MCAR), sometimes called uniform non-response, is of no interest in this study. This is because the propensity for any given data to miss does not depend on any observed or unobserved measurements. Therefore, the complete data records can be considered to be a random sub-sample of the original data (Allison, 2001; Heitjan & Basu, 1996). Data missing at random (MAR) mechanism indicates that missingness is unrelated to the missing values themselves but is related to another variable. Hence, when conditioned on the variable, the missingness is completely random. MAR data are ignorable, thereby eliminating the need to model the missing data mechanism. Maximum likelihood inferences are valid under MAR (Allison, 2001; Donald B. Rubin, 1976). If missingness is non-ignorable, hence related to

unobserved variables, then the missing data mechanism needs to be modeled to get valid estimates of model parameters. One such situation occurs when missingness depends on unobserved random intercepts or random slopes can be made (Park, et al., 2002). In this case, pattern mixture models can be applied (Little, 1993, 1995). A novel idea in this research is to incorporate the pattern of missing data into propensity scores that will be utilized in the statistical models.

First, the MAR missing data mechanism is considered. In this case, missingness was assumed to depend on gender. The mean parameter estimates computed from the 5000 simulated complete dataset, despite the application of amplified random effects, were similar to those of the bias literacy CRT. Out of the 5000 instances of complete datasets, only one instance was used to generate the gender attrition dataset (Table 7). However, some parameter estimates of the data instance selected to generate missingness moderately differed from the mean simulated estimates of the 5000 complete datasets. The variation in parameter estimates was not unique to a small set of instances. Data instances varied a lot. The amplified random effects contributed to this variation. The process of creating gender attrition using an instance of complete dataset was a Bernoulli trial with the probability of men and of women missing at any given time point (baseline, 3-Day and 3-Month) and in any given experimental group being 55% and 40% respectively. The mean parameter estimates computed from the 5000 simulated gender attrition datasets are similar to the complete data used to generate the attrition (Table 7). This demonstrates that MAR data missingness is ignorable when using maximum likelihood estimates. Furthermore, there were 22 fewer responses on average after attrition simulations than there were in the bias literacy

CRT. Hence, the magnitude of missingness was similar to the magnitude of missingness in the bias literacy CRT.

Table 7: Comparison of Parameter Estimates with Gender Attrition Parameter Estimates

Description of Parameter Estimators	Bias Literacy Dataset	Complete Dataset 5000 Simulations†	Complete Dataset One Instance	Gender Attrition 5000 Simulations†
Intercept	5.3912	5.3803	5.5261	5.5262
Group (Experimental Group)	-0.0391	-0.0348	-0.0782	-0.0754
Time Period (3-Day)	0.0673	0.0625	0.0617	0.0625
Time Period (3-Month)	0.1072	0.1031	0.1377	0.1345
Gender (Female)	0.1543	0.1460	0.0347	0.0356
Interaction (3-Day and Exp. Group) ††	0.1659	0.1714	0.1717	0.1700
Interaction (3-Month and Exp. Group) ††	0.2055	0.2047	0.1928	0.2019
Average No. of Observations	1497	2895	2895	1475

† The simulation columns provide the means of the estimated parameters.
†† The treatment effect is the mean difference at follow-up minus the mean difference at baseline between the intervention and control group.

Table 8: Comparison of Standard Errors with Gender Attrition Standard Errors

Description of Parameter Estimators	Bias Literacy Dataset	Complete Dataset 5000† Simulations	Complete Dataset One Instance	Gender Attrition 5000† Simulations
Intercept	0.0554	0.6898	0.7378	0.7411
Group (Experimental Group)	0.0643	0.7054	0.7532	0.7571
Time Period (3-Day)	0.0539	0.0333	0.0334	0.0544
Time Period (3-Month)	0.0562	0.0333	0.0334	0.0544
Gender (Female)	0.0518	0.0489	0.0493	0.0565
Interaction (3-Day and Exp. Group)	0.0742	0.0464	0.0467	0.0757
Interaction (3-Month and Exp. Group)	0.0788	0.0464	0.0467	0.0757
Average No. of Observations	1497	2895	2895	1475

† The simulation columns provide the means of the estimated standard errors.

The standard errors of the complete dataset and the gender attrition dataset are more closely aligned with each other than with the bias literacy analysis. This is because the bias literacy data analysis has a different random effects structure.

Second, the non-ignorable missing mechanism is demonstrated. Like in the MAR missing mechanism, the parameters for the non-ignorable missingness were calibrated to generate a similar magnitude of missingness as in the bias literacy CRT. Three cases of non-ignorable data attrition were constructed: attrition based on random intercept (RI); attrition based random effects but with differing magnitude of missingness based on the experimental group (RIE); and attrition based random effects, experimental group and follow-up time (RIET). In all the three cases, a fitted logistic function is used to generate probabilities that are used to create attrition. The parameters in the logistic function are used to determine the different types of attrition (3.2):

$$\pi = \frac{1}{\left(1 + e^{(-X\beta)}\right)}. \tag{3.2}$$

In this case, π is the vector of probabilities of attrition and X is the model matrix and β is the vector of desired attrition co-efficient. The generated probabilities are then used in a Bernoulli trial to generate attrition. The same instance of compete data that was used to generate gender attrition was also used to generate all the different cases of non-ignorable attrition. In the first case, in which attrition is based on random intercept (RI), missingness does not depend on gender; hence, gender is similarly distributed across experimental groups and time (Figure 4).

Figure 4: Non-Respondents by across time Periods

The means of the parameter estimates from the simulations indicated a 5.5% increase and 4.3% increase in the treatment effect in 3-Day time period and 3-Month time period respectively when compared instance of the complete dataset from which attrition was generated (Table 9). While missingness in this case is non-ignorable, it similarly affects the two groups making it equivalent to randomly picking participants. On average there were 53 fewer observations in the simulated datasets than there were in the bias literacy CRT.

Table 9: Impact of Random Intercept (RI) Attrition on Parameter Estimates

Description of Parameter Estimators	Complete Dataset One Instance Estimates	RI Attrition Simulations (5000) Mean Estimates
Intercept	5.5261	5.5772
Group (Experimental Group)	-0.0782	-0.0676
Time Period (3-Day)	0.0617	0.0555
Time Period (3-Month)	0.1377	0.1322
Gender (Female)	0.0347	0.0419
Interaction (3-Day and Exp. Group)	0.1717	0.1812
Interaction (3-Month and Exp. Group)	0.1928	0.2011
Average No. of Observations	2895	1444

† mean difference at follow-up minus the baseline difference between the intervention and control group

The second case of non-ignorable mechanism is missingness that depends on random effects with the magnitude of missingness differing based on the experimental group to which the person belongs (RIE). It was assumed that individuals in the intervention group were more likely than individuals in the control to continue with the study during the follow-up and hence were assigned a lower rate of attrition. The attrition process was simulated 5000 times. The mean parameter estimates for the 3-Day and 3-Month treatment effect were overestimated by 15.7% and 9.7% respectively (Table 10). Based on the simulations, there were 53 fewer observations after the attrition than there were in the bias literacy CRT. An instance of the RIE dataset indicates that attrition in the intervention group was slightly lower than that of the control group (Figure 5).

Table 10: Impact of Random Intercept By Experimental Group (RIE) Attrition

Description of Parameter Estimators	Complete Dataset One Instance Estimates	RIE Attrition Simulations (5000) Mean Estimates
Intercept	5.5261	5.7082
Group (Experimental Group)	-0.0782	-0.0989
Time Period (3-Day)	0.0617	0.0406
Time Period (3-Month)	0.1377	0.1248
Gender (Female)	0.0347	0.0670
Interaction (3-Day and Exp. Group)	0.1717†	0.1987
Interaction (3-Month and Exp. Group)	0.1928†	0.2115
Average No. of Observations	2895	1444

† mean difference at follow-up minus the baseline difference between the intervention and control group

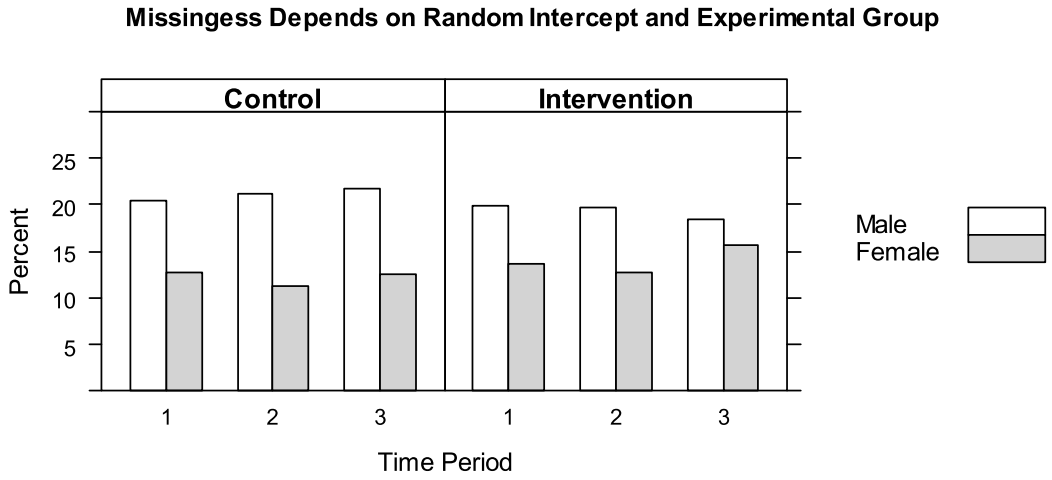

Figure 5: Non-Respondents across time Periods

In the third case, missingness depends on random effects/intercept, experimental group and time (RIET). In this case, there was an average of 28 more observations for the simulated datasets after attrition than there was in the bias literacy CRT. The 3-Day treatment effect was overestimated by 19.1% when compared to the instance of the complete dataset (Table 11). On the other hand, the 3-Month treatment effect was underestimated by 9.4%. In this case, the differential attrition across time and experimental group is more evident than in the RIE attrition (Figure 6).

Table 11: Impact of Random Intercept By Experimental Group (RIET) Attrition

Description of Parameter Estimators	Complete Dataset One Instance Estimates	RIET Attrition Simulations (5000) Mean Estimates
Intercept	5.5261	5.7064
Group (Experimental Group)	-0.0782	-0.1234
Time Period (3-Day)	0.0617	0.0405
Time Period (3-Month)	0.1377	0.1246
Gender (Female)	0.0347	0.0692
Interaction (3-Day and Exp. Group) †	0.1717	0.2045
Interaction (3-Month and Exp. Group) †	0.1928	0.1747
Average No. of Observations	2895	1525

† mean difference at follow-up minus the baseline difference between the intervention and control group

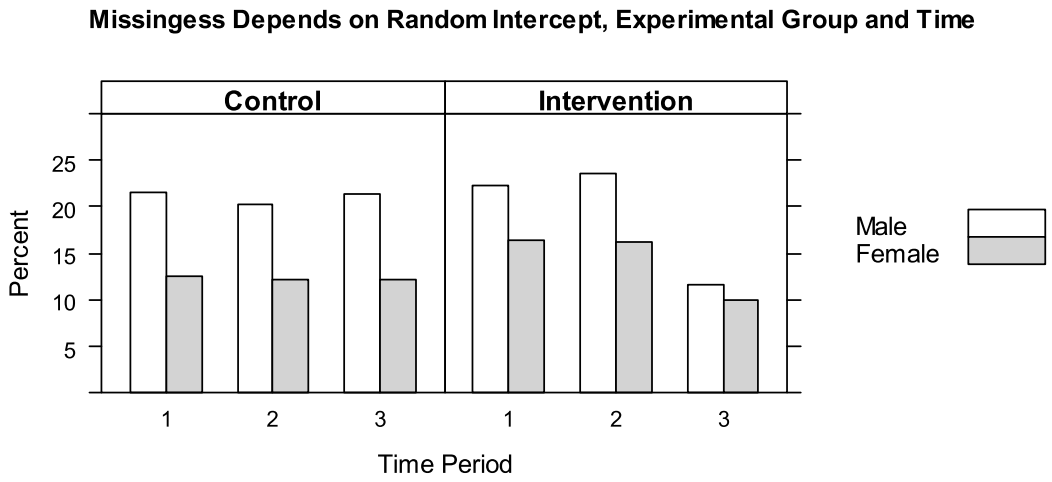

Figure 6: Non-Respondents across time Periods

3.2.2 Propensity Score Estimation

Propensity scores were estimated using logistic regression models. The objective was to estimate, for each individual, the probability of being assigned to the intervention group given cluster level covariates, given gender as an individual level covariate, and given either a random intercept or missing data attributes. There were two types of missing data attributes explored in this research. The first is the frequency of observations across the follow-up time periods for each individual after attrition. The second attribute is the missing data pattern. The patterns are formed by the sequence of indicators in the follow-up time periods (Baseline, 3-Day, 3-month) that indicate missingness when the indicator is 0 or indicate presence of a response when the indictor is 1. Cluster level covariates were included in propensity score estimation because they were used as matching variables to mitigate imbalance of baseline covariates in the bias literacy CRT. There were six cluster level variables used in estimating propensity scores. The first was the school in which the department belongs. The schools in the bias literacy CRT included: College of Agricultural and Life

Sciences, College of Letters and Science, College of Engineering, School of Pharmacy, School of Medicine and Public Health, and School of Veterinary Medicine. The second was the division in which the majority of faculty members were tenured. The divisions included in the study consisted of: biological sciences, physical sciences, social studies and arts and humanities. The third cluster level variable was the size of the department based on the number of faculty (small = 0-13 faculty, medium = 14-23 faculty, large= 24+ faculty). The fourth variable was the percent of junior faculty in the department; for this variable, low was defined as less than 20%, medium as 20% to less than 40% and high as 40% or greater. The fifth cluster level variable was the percent of total faculty who are in the clinical health sciences (CHS) track as opposed to tenure track, where 0 represented departments with no CHS faculty, 1 represented department with less than 50% CHS faculty, 2 represented departments with 50% or more to less than 75% CHS faculty, and 3 represented departments with 75% or more CHS faculty. The sixth cluster level variable was the percent female in the department, where low was less than 10% for physical sciences and less than 25% for other divisions, medium was 10% to less than 20% for physical sciences and 25% to less than 40% for other divisions, and high was equal to or greater than 20% for physical sciences and equal to or greater than 40% for other divisions.

Logistic regression models were used to estimate propensity scores. The outcome variable was binary with the control group assigned to 0 and the intervention group assigned to 1. The logistic response function (3.3) is monotonic and sigmoidal in shape with respect to $X'\beta = \beta_0 + \beta_1 X_1 + ... + \beta_{p-1} X_{p-1}$:

$$E(Y) = \frac{\exp(X'\beta)}{1+\exp(X'\beta)} \quad \text{Where,} \quad \underset{p\times 1}{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{P-1} \end{bmatrix} \quad \underset{p\times 1}{X} = \begin{bmatrix} 1 \\ X_1 \\ \vdots \\ X_{P-1} \end{bmatrix}. \tag{3.3}$$

All model selection used for the logic regression to estimate propensity score was carried out using the Akaike information criterion (AIC) in a bi-directional stepwise algorithm, which measures relative quality of a statistical model while balancing between the goodness of fit and complexity of the model. A full model, prior to applying the AIC stepwise model selection algorithm, consisted of a set of three way interactions in which each one characteristic of missingness, interacted with each of the cluster level variables and with the only individual level covariate which is gender (Table 12).

Table 12: Missingness Characteristics and Cluster and Individual Variables Used for PS Estimation

Characteristics of Random Intercept Missingness	Department/Cluster Level Variables	Individual Level Variable
Random Intercept or Sum of Observations or Attrition Pattern	School Division Size Percent Junior Faculty Percent Faculty in Clinical Health Sciences Percent Women Faculty	Gender

There were three types of attrition datasets used for the research: dataset RI in which missingness depended on random intercept only; dataset RIE in which missingness depended on random effects with different magnitude of missingness based on the experimental group of the participants; dataset RIET in which missingness depended on random intercept, experimental group and time. Two types of approaches were considered. In the first approach, all individuals regardless of attrition outcome were

included in propensity score calculations for the attrition datasets RI, RIE, RIET. However, in the second approach, individuals with missing response outcomes at all three time points (baseline, 3-Day and 3-Month) were excluded from the propensity score model adjustment analysis (Figure 7).

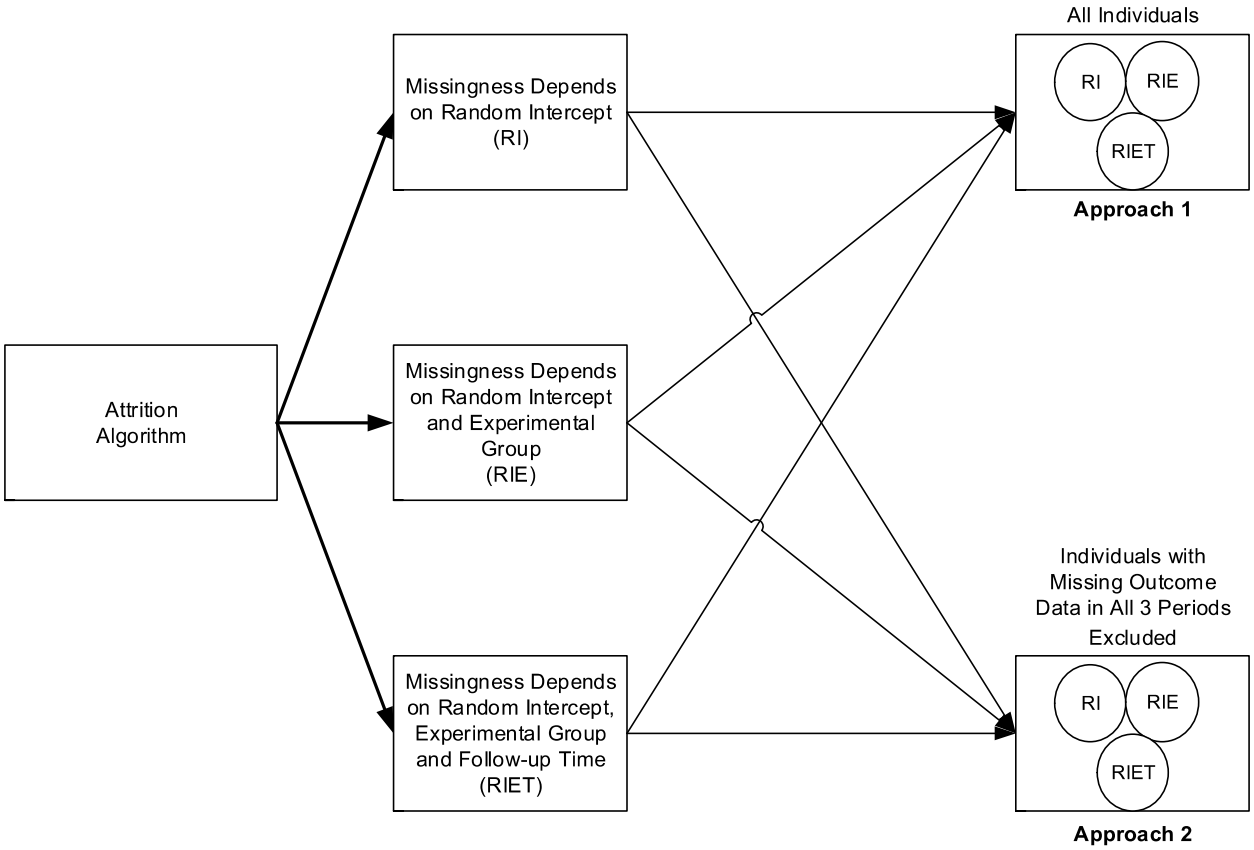

Figure 7: Attrition Datasets used for Propensity Score Estimation

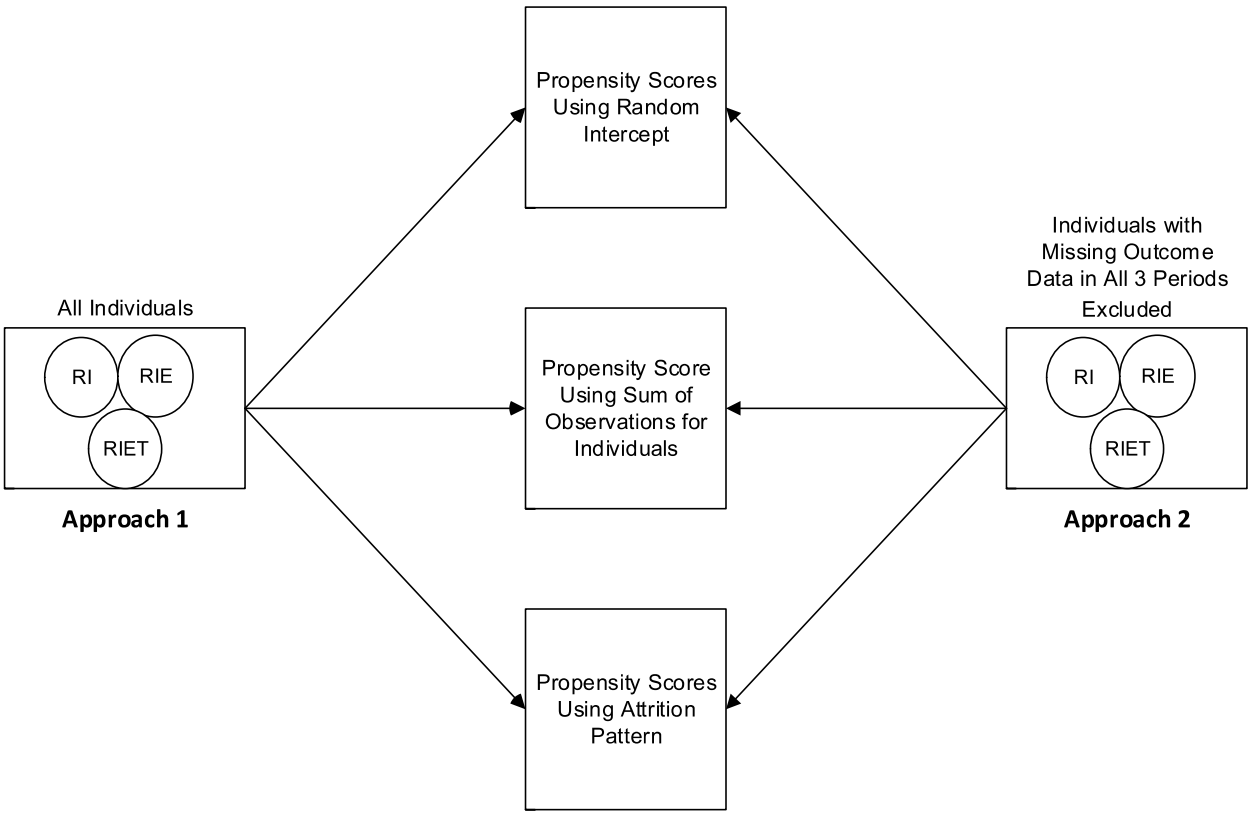

Figure 8: Propensity score estimated using different attrition attributes in three types of attrition datasets.

In the first approach, all individuals are included in estimation of propensity scores in all three attrition datasets (RI, RIE, and RIET). Propensity scores based on random intercepts, the number of observations after attrition, and missingness patterns are estimated for each attrition data set (Figure 8). Direct modeling of random effects is a validity test or best case scenario which may not be reproducible in practice. In the second approach, individuals with missing outcome data in the baseline and two follow-up periods are omitted from the propensity calculation.

3.2.3 Data Analysis

The data was analyzed using R software for statistical computing (Team R Core, 2014) using the lme4 package for linear mixed effects models (D. Bates, et al., 2012; D. M. Bates, 2010). The same models were fit to all datasets. The response variable is the

efficacy score. Two types of models are considered for the analysis: models in which the propensity scores are adjusted as a covariate and models in which propensity score is included as an interaction term. The efficacy score or the outcome variable is y_{ijkt} where $i = 0,1$, $j = 1,...,m$, $k = 1,...,n_j$, and $t = 1,2,3$. This represents the i^{th} experimental group, j^{th} cluster with the k^{th} person at the t^{th} follow-up time. The gender of the k^{th} person is represented by an indicator variable I_{g_k} in which the k^{th} person's gender (g) is 0 if male and 1 if female. The propensity score x_{ijk} is either modeled to interact with the experimental group indicator variable and the follow-up time (3.4) or is modeled as a covariate that is adjusted in the model (3.5).

Model with Propensity Score as an Interaction:

$$\begin{aligned} y_{ijkt} = {} & \beta_0 + \beta_1 \cdot I_{i=1} + \beta_2 \cdot I_{t=t_2} + \beta_3 \cdot I_{t=t_3} + \beta_4 x_{ijk} + \beta_5 \cdot I_{g_k=1} \\ & + \beta_6 \cdot I_{i=1,t=t_2} + \beta_7 \cdot I_{i=1,t=t_3} \\ & + \beta_8 x_{ijk} \cdot I_{i=1} + \beta_9 x_{ijk} \cdot I_{t=t_2} + \beta_{10} x_{ijk} \cdot I_{t=t_3} \\ & + \beta_{11} x_{ijk} \cdot I_{i=1,t=t_2} + \beta_{12} x_{ijk} \cdot I_{i=1,t=t_3} \\ & + \gamma_{ij} + \lambda_{1j} + \omega_{ijk} + \varepsilon_{ijkt}. \end{aligned} \tag{3.4}$$

Model with Propensity Score as a Covariate:

$$\begin{aligned} y_{ijkt} = {} & \beta_0 + \beta_1 \cdot I_{i=1} + \beta_2 \cdot I_{t=t_2} + \beta_3 \cdot I_{t=t_3} + \beta_4 x_{ijk} + \beta_5 \cdot I_{g_k=1} \\ & + \beta_6 \cdot I_{i=1,t=t_2} + \beta_7 \cdot I_{i=1,t=t_3} \\ & + \gamma_{ij} + \lambda_{1j} + \omega_{ijk} + \varepsilon_{ijkt}. \end{aligned} \tag{3.5}$$

All propensity scores are centered on the overall average propensity score. The different types of propensity scores will be applied in models with the outcome variables

with data missingness corresponding to the attrition datasets used. Bias percent will be used to assess the degree of correction. Bias of an estimate, $\hat{\theta}$, is defined as follows:

$$bias\left(\hat{\theta}\right) = E\left(\hat{\theta}\right) - \theta \quad or \quad Percent \;\; bias\left(\hat{\theta}\right) = \frac{E\left(\hat{\theta}\right) - \theta}{\theta} x100 \tag{3.7}$$

Estimator unbiased if: $E\left(\hat{\theta}\right) = \theta.$

In this case, θ represents the truth. In this case, mean θ is the treatment effect parameter estimates of the model simulations of the full dataset (Table 7). The parameter $E\left(\hat{\theta}\right)$ represents the treatment effect parameter estimates using attrition data instances. The treatment effect parameter estimate is the coefficient estimated from the interaction between the experimental group and the follow-up time. The treatment effect coefficient is the mean difference minus the baseline difference between the intervention and control group at 3-Day and at 3-Month time points. The random components of the simulations are determined by the following random effects: γ_{ij} which represents between-cluster variability; λ_{1j} which represents variability between the interaction cluster and experimental group to account for the variability in the treatment effect; ω_{ijk} which represents the variability between persons; and ε_{ijkt} which represents random error.

$$\gamma_{ij} \sim N\left(0, \sigma^2_{b_1}\right), \quad \lambda_{1j} \sim N\left(0, \sigma^2_{b_2}\right), \quad \omega_{ijk} \sim N\left(0, \sigma^2_{b_3}\right), \quad \text{and } \varepsilon_{ijkt} \sim N\left(0, \sigma^2\right) \tag{3.8}$$

The methods described in this chapter will help guide the generation of dataset with different characteristics, the estimation of 3 types of propensity scores and the application of such scores to models for data analysis.

CHAPTER 4
RESULTS

This chapter begins with a discussion of the characteristics of estimated propensity scores. The discussion is followed by a detailed review of the impact of using propensity scores on bias. The chapter ends with the application of propensity score methods to the bias literacy cluster randomized trial.

4.1 Propensity Score Analysis

Propensity scores calculated using characteristics of missingness alone or characteristics of missingness in combination with cluster and individual level variables are examined for the degree of linearity with propensity scores calculated using random effects (Table 12). This is to confirm previous results indicating that missingness depending on random effects can be captured by missing data patterns (Park, et al., 2002). The results show that direct modeling of propensity scores on random effects is a validity check as random effects cannot be assumed to be correctly estimable in practice.

Propensity scores estimated using the sum of observations, under Approach 1 (Figure 8) for all of the attrition datasets (RI, RIE, and RIET) are examined. Under Approach 1, all individuals including those with missing outcome data in all three periods are included in the propensity score calculation. All propensity scores are centered about the average propensity score of all participants.

In general, propensity scores estimated using the numbers of observations are linearly related to the propensity scores estimated using random effects regardless of the type of attrition dataset (Figure 9). This is because the relative weights of the propensity scores are the same. The linearity can be further examined using graphs partitioned by attrition datasets (Figure A1.1 to A1.3). There was a strong linear relationship between propensity scores estimated from random effects and those using the number of observations across the attrition datasets. The coefficient of determination for the RI, RIE, and RIET attrition dataset was 0.84, 0.80, and 0.76 respectively. The coefficient of determination indicates the proportion of the total variation that can be attributed to the linear relationship between propensity scores as opposed to random variation. There was some random variation around the scatter plot LOWESS curve, a locally-weighted polynomial regression (Figure 10). This is because there are only three observations per person. Differential attrition associated with RIE and RIET datasets further contributed to the increased variation around the LOWESS curve. The scatter plot LOWESS curve indicates minimal differences in the linearity characteristic.

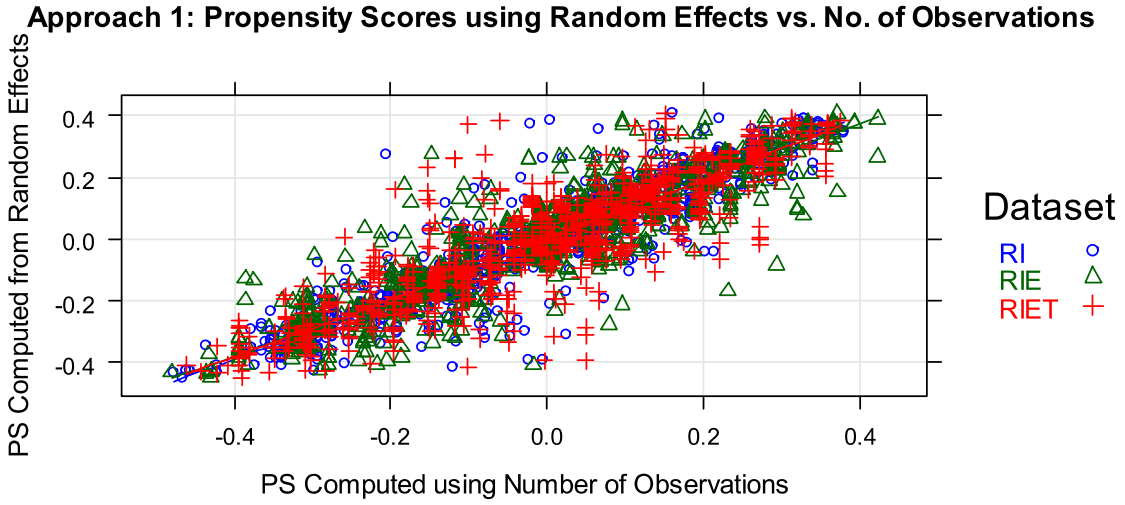

Figure 9: All individual outcomes are included in propensity score estimation regardless of attrition.

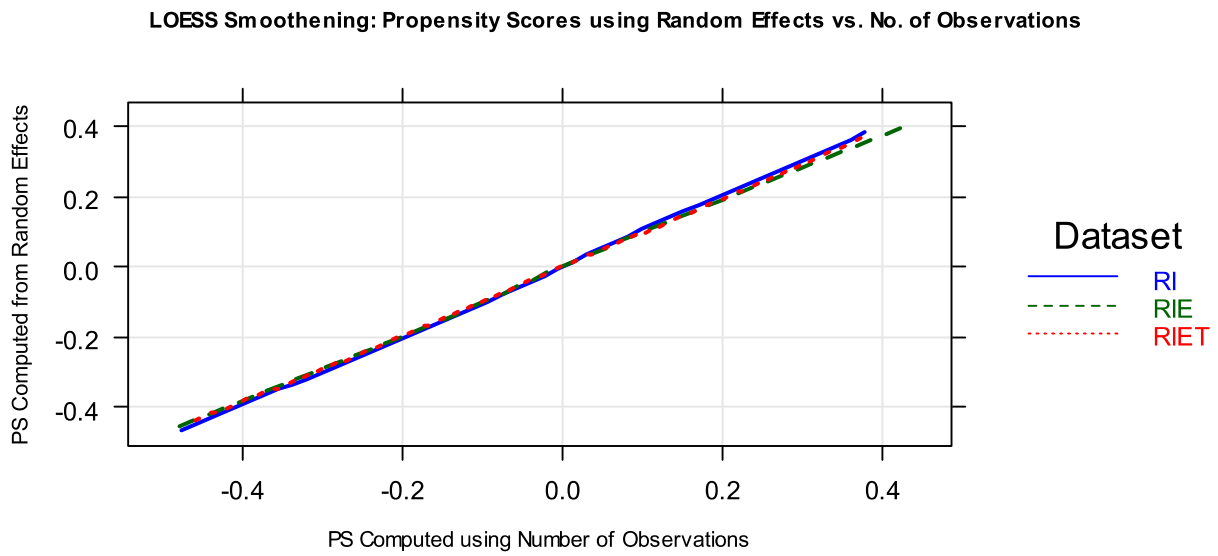

Figure 10: All individual outcomes are included in propensity score estimation regardless of attrition.

Propensity scores estimated from the missing data pattern exhibited the same characteristics as those estimated from the number of observations. There seems to be an equivalent amount of variance around the average propensity scores estimated using missing data patterns as there was in propensity scores estimated using the number of observations (Figure 11). The linear relationship between propensity scores estimated using missing data patterns and propensity scores estimated using random effects was strong as indicated by coefficient of determination of 0.85, 0.81, and 0.81for the RI, RIE, and RIET attrition datasets respectively (Figures A1.4 to A1.6). The scatter plot LOWESS curve yielded similar results as those of the number of observations (Figure 12). These results indicate that under Approach 1, propensity scores estimated from the number of observations and missing data pattern are linearly related to the random effects.

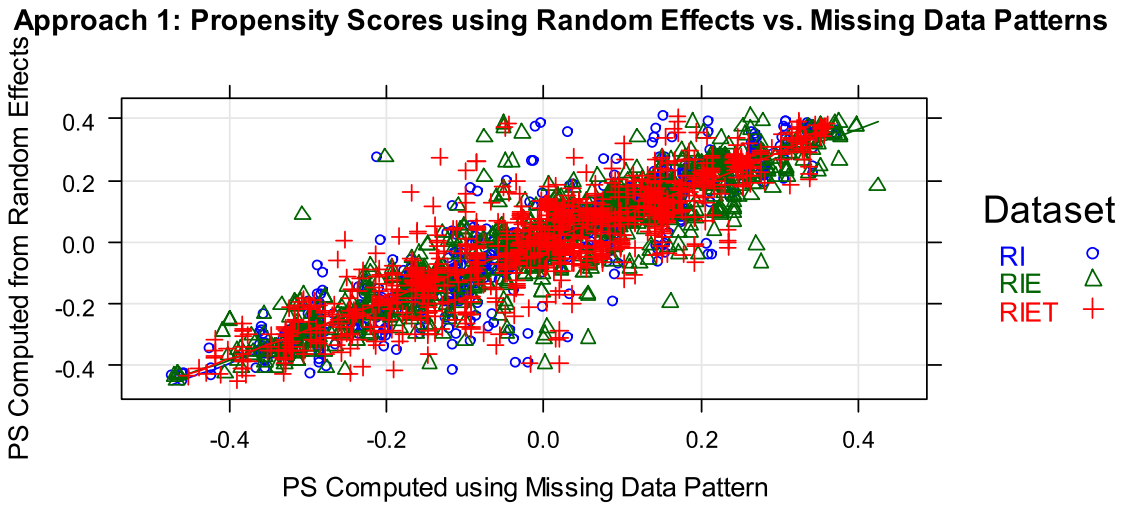

Figure 11: All individuals regardless of attrition outcome are included in propensity score estimation

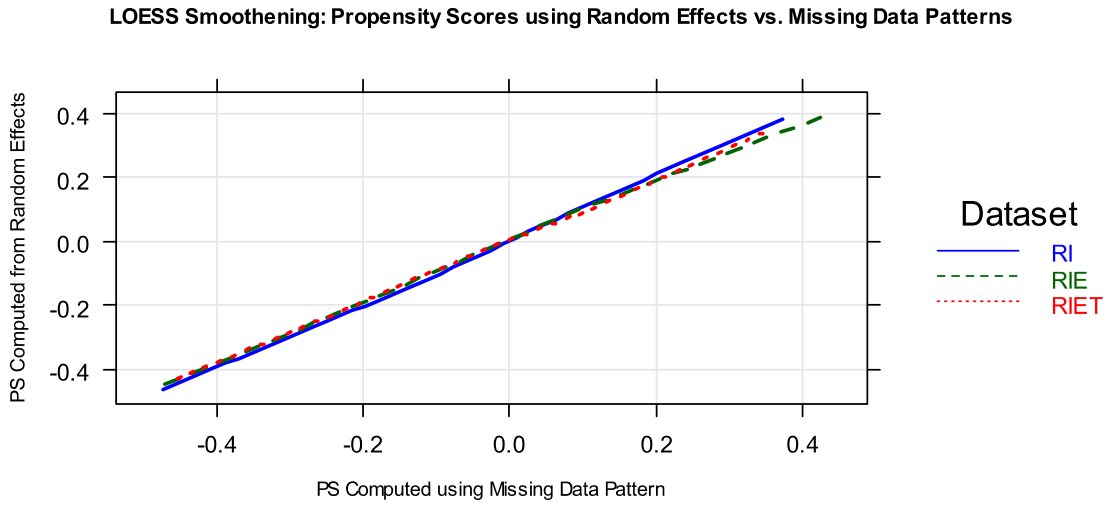

Figure 12: All individuals regardless of attrition outcome are included in propensity score estimation

Propensity scores estimated under both Approaches 1 and 2 (Figure 8) are examined next. Propensity scores that are estimated using random effects under both approaches are first examined. Based on the results, for all three types of attrition datasets (RI, RIE, RIET), propensity scores under Approach 1 are linearly related to those under Approach

2 with a coefficient of determination 0.87, 0.80 and 0.90 respectively (Figures 13, 14, and 15).

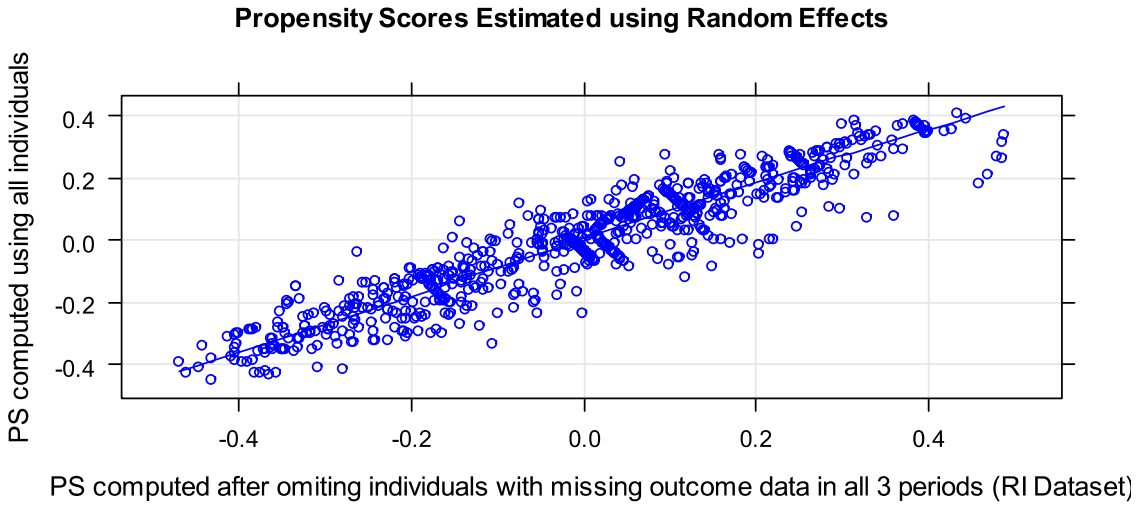

Figure 13: There is a linear relationship between propensity scores estimated under Approach 1 and those estimated under Approach 2

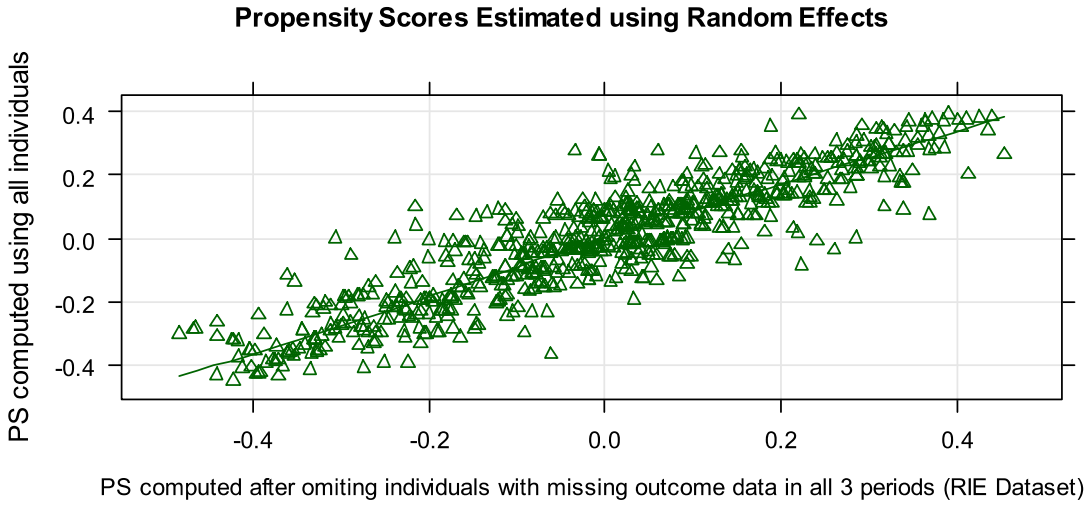

Figure 14: There is a linear relationship between propensity scores estimated under Approach 1 and those estimated under Approach 2

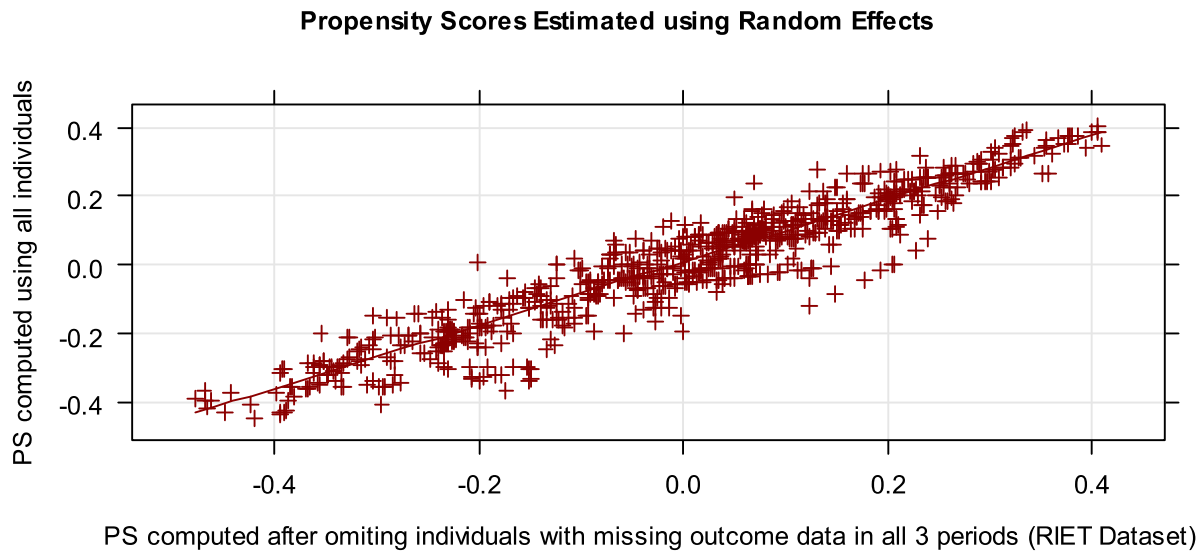

Figure 15: There is a linear relationship between propensity scores estimated under Approach 1 and those estimated under Approach 2

Next, the propensity scores calculated using the number of observations for the attrition datasets under Approach 1 are compared to those calculated under Approach 2. When using the RI attrition data, the propensity scores estimated using all individuals (Approach 1) are linearly related to those estimated after omitting individuals with missing outcome data in all 3 periods (Approach 2). There is minimal variation about the LOWESS curve (Figure 16). Minimal variation is confirmed by a coefficient of determination of 0.96.

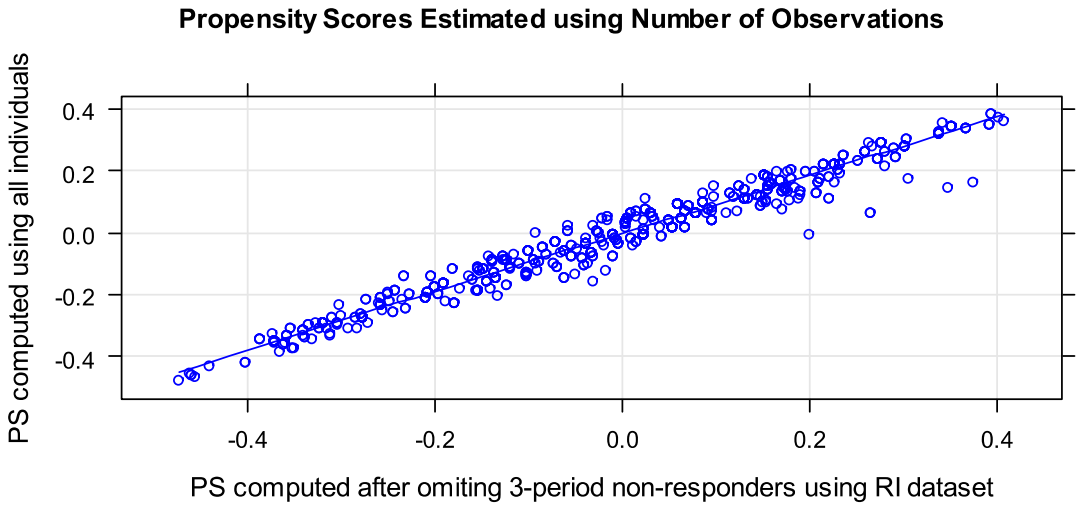

Figure 16: There is a linear relationship between propensity scores estimated under Approach 1 and those estimated under Approach 2

In the case of the RIE and RIET datasets, while there is a linear relationship between propensity scores estimated using the sum of observations under Approach 1 and those estimated under Approach 2, there is more variation about the LOWESS curve (Figures 17 and 18). The variation may be attributed to the level of bias introduced by the two non-ignorable datasets. The coefficient of determination was 0.73 and 0.92 respectively.

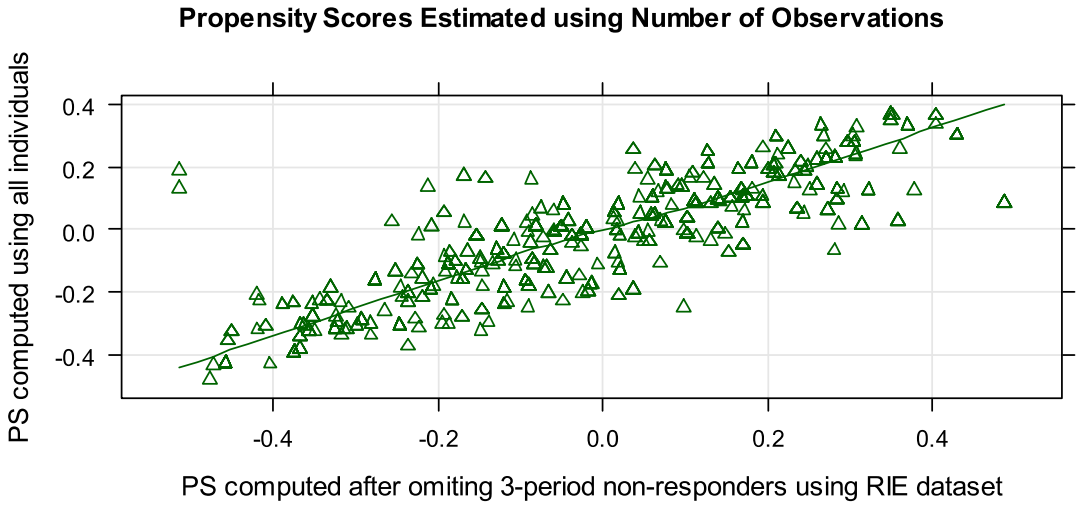

Figure 17: There is a linear relationship between propensity scores estimated under Approach 1 and those estimated under Approach 2

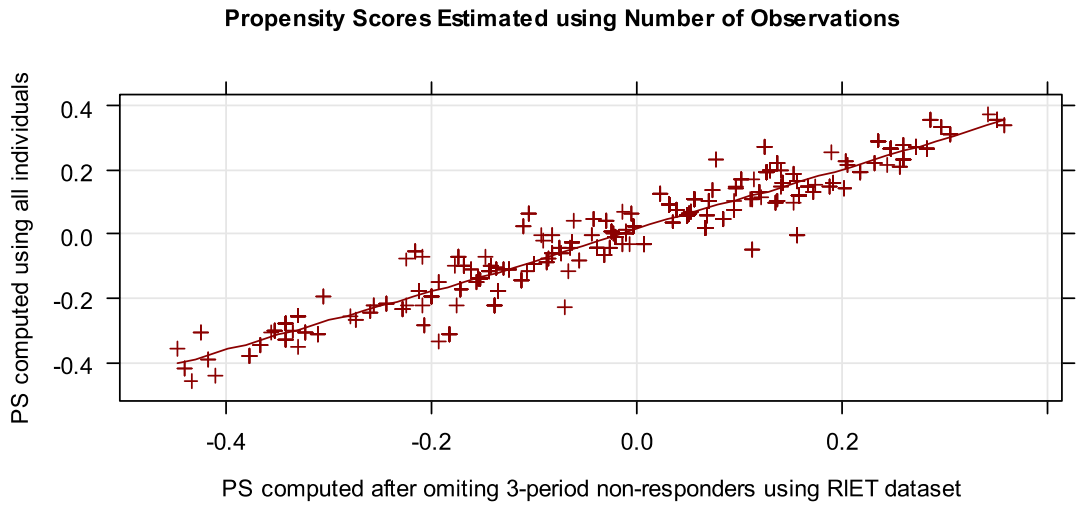

Figure 18: There is a linear relationship between propensity scores estimated under Approach 1 and those estimated under Approach 2

Similar characteristics were found when examining the linear relationship between propensity scores calculated using the missing data pattern under Approach 1 and those calculated under Approach 2 (Figures 19, 20 and 21). The coefficient of determination corresponding with the figures was 0.98, 0.83, and 0.86 respectively. The results indicate that non-ignorable missingness can be captured by both missing data patterns and number of observations. Furthermore, differential attrition, as created in RIE and RIET attrition datasets, has an impact on the variability of estimated propensity scores about the LOWESS curve while still maintaining a linear relationship with random effects.

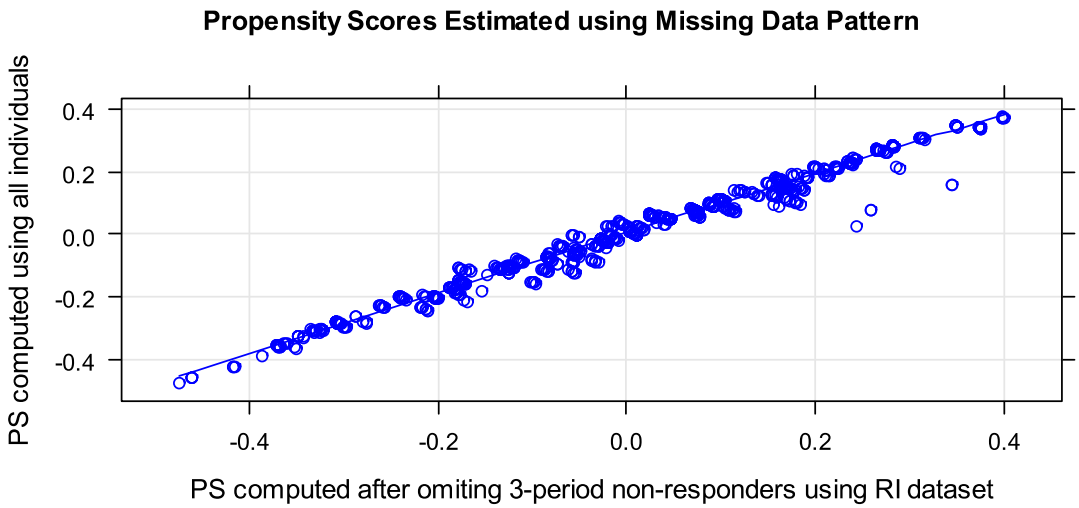

Figure 19: There is a linear relationship between propensity scores estimated under approach 1 and those estimated under approach 2

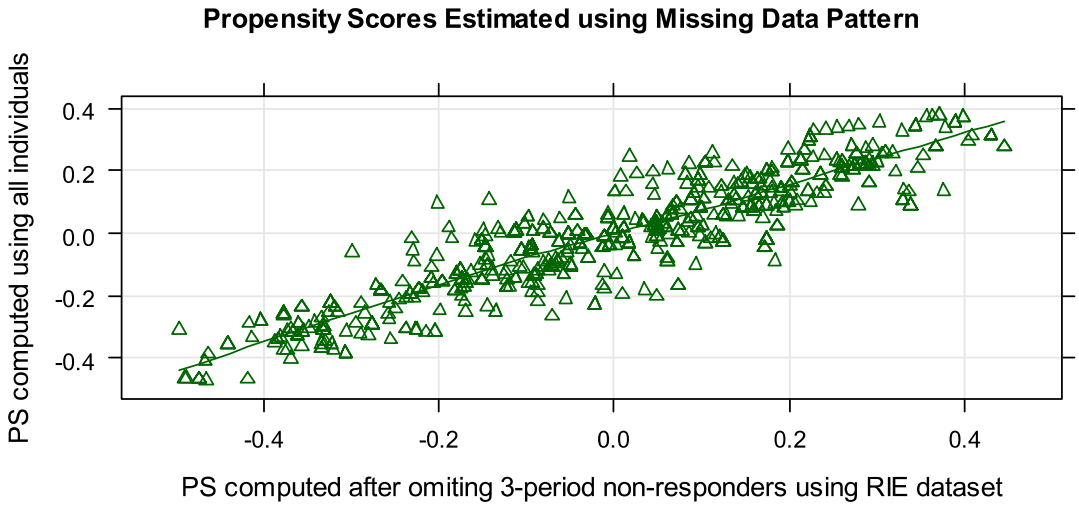

Figure 20: There is a linear relationship between propensity scores estimated under approach 1 and those estimated under approach 2

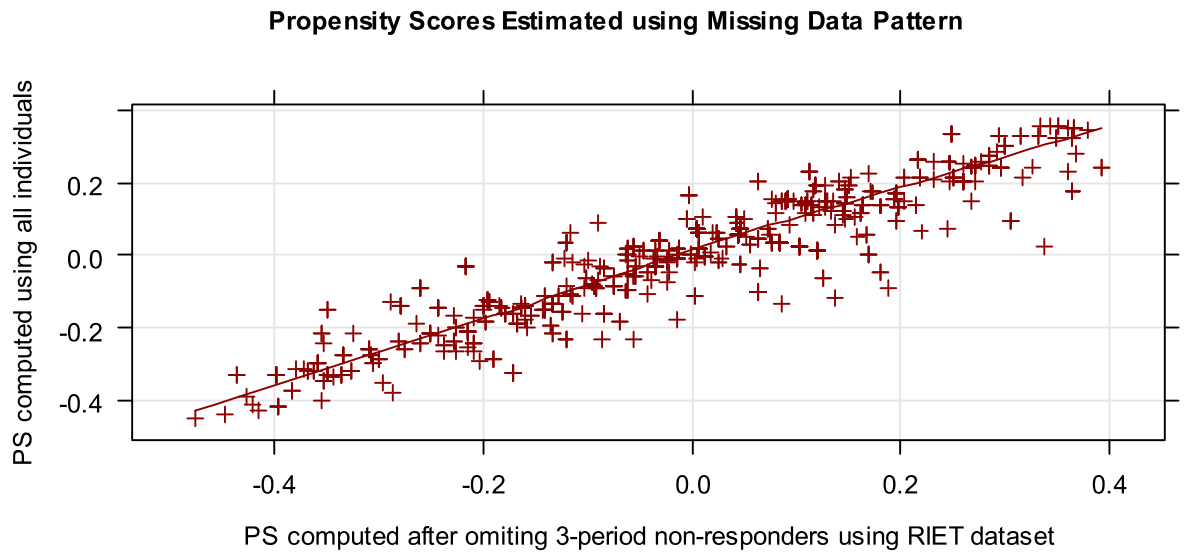

Figure 21: There is a linear relationship between propensity scores estimated under approach 1 and those estimated under approach 2

4.2 Impact of Propensity Scores on Data Analysis

The application of propensity scores will be examined in this section. Propensity scores were applied to the models, under Approach 1 and Approach 2 (Figure 8), as either an interaction as in (3.4), or as a covariate as in (3.5). First, the application of propensity scores under Approach 1, in which all individuals are included in the estimation of propensity scores, will be examined. The bias, given by (3.7), in the treatment effect for the unadjusted, covariate adjusted and interaction adjusted models is examined. The treatment effect coefficient is the mean difference between the control and experimental group minus the baseline mean difference between the intervention and control group at 3-Day and at 3-Month time points. The unadjusted, covariate adjusted and interaction adjusted models are labeled as 1-U, 2-C, and 3-I respectively in the relevant figures provided. In this analysis, no bias is considered to be 0%. A bias threshold of $\pm 10\%$ is considered to be acceptable.

4.2.1 Performance of Propensity Scores under Approach 1

Propensity Scores Computed using Random Intercept at 3-Day

Under Approach 1 (Figure 8), for the 3-Day follow-up period, the bias in the mean treatment effect estimates in the unadjusted models using the RI attrition dataset was 2.0%, indicating a minimal overestimation of the treatment effect. The covariate adjusted models reduced the bias on average from 2.0% to -0.4%. On the other hand, adding the propensity score to the models as an interaction term with the treatment effect changed the bias from an overestimation of the mean treatment effect of 2.0% to an underestimation of the mean treatment effect of 5.2% (Table 13). In all cases, bias is within the acceptable $\pm 10\%$ threshold. The bias is negligible in this instance because attrition is based on random effects only. As is the case in the RI attrition dataset, bias affects both of the experimental and control groups similarly.

In the case of the RIE attrition dataset, the bias in mean treatment effects for unadjusted models in the 3-Day follow-up period is 15.7%. The bias in the RIE attrition dataset was significantly bigger than the bias in the RI attrition dataset. The interaction adjusted models reduced the bias from 15.7% to 1.5%, while the covariate adjusted model reduced the bias from 15.7% to 10.5% (Table 14).

The bias mean treatment effect from unadjusted models in the RIET attrition dataset was 19.0%. Like the RIE attrition dataset, the interaction adjusted model reduced bias more efficiently to 4.4% than did the covariate adjusted model, which reduced bias to 14.6% (Table 15).

Table 13: Approach 1 - Impact of Propensity Scores using Random Intercept in the RI Attrition Dataset

Description of Parameter Estimators	Complete Dataset 5000 Simulations	Coefficient Means Unadjusted Models	Coefficient Means Interaction Adjusted	Coefficient Means Covariate Adjusted
Intercept	5.3803	5.5691	5.4632	5.4647
Group (Experimental Group)	-0.0348	-0.0514	0.0031	0.0018
Time Period (3-Day)	0.0625	0.0508	0.0590	0.0521
Time Period (3-Month)	0.1031	0.1345	0.1415	0.1369
Propensity Score	*	*	-25.7147	-25.5943
Gender (Female)	0.1460	0.0451	0.1332	0.1332
Treatment Effect (3-Day and Exp. Group)	0.1714	0.1749	0.1626	0.1707
Treatment Effect (3-Month and Exp. Group)	0.2047	0.1973	0.1843	0.1905

Table 14: Approach 1 – Impact of Propensity Scores using Random Intercept in the RIE Attrition Dataset

Description of Parameter Estimators	Complete Dataset 5000 Simulations	Coefficient Means Unadjusted Models	Coefficient Means Interaction Adjusted	Coefficient Means Covariate Adjusted
Intercept	5.3803	5.7067	5.4719	5.4673
Group (Experimental Group)	-0.0348	-0.1000	-0.0171	-0.0075
Time Period (3-Day)	0.0625	0.0449	0.0620	0.0506
Time Period (3-Month)	0.1031	0.1344	0.1448	0.1315
Propensity Score	*	*	-25.1627	-25.7085
Gender (Female)	0.1460	0.0701	0.1396	0.1390
Treatment Effect (3-Day and Exp. Group)	0.1714	0.1983	0.1741	0.1895
Treatment Effect (3-Month and Exp. Group)	0.2047	0.2090	0.1890	0.2061

Table 15: Approach 1 – Impact of Propensity Scores using Random Intercept in the RIET Dataset

Description of Parameter Estimators	Complete Dataset 5000 Simulations	Coefficient Means Unadjusted Models	Coefficient Means Interaction Adjusted	Coefficient Means Covariate Adjusted
Intercept	5.3803	5.7072	5.4704	5.4670
Group (Experimental Group)	-0.0348	-0.1205	-0.0185	-0.0097
Time Period (3-Day)	0.0625	0.0409	0.0603	0.0422
Time Period (3-Month)	0.1031	0.1284	0.1392	0.1240
Propensity Score	*	*	-25.1965	-25.6591
Gender (Female)	0.1460	0.0618	0.1391	0.1390
Treatment Effect (3-Day and Exp. Group)	0.1714	0.2040	0.1790	0.1965
Treatment Effect (3-Month and Exp. Group)	0.2047	0.1712	0.1987	0.2125

Propensity Scores Computed using Sum of Observations at 3-Day

Similar to propensity scores estimated using random intercept and the RI attrition dataset, propensity scores estimated using the sum of observations and the RI attrition dataset showed little bias across the different types of models. The bias remained within the $\pm 10\%$ threshold. The mean treatment effect for the unadjusted models showed a bias of 2.0%, while that for interaction adjusted models had a bias of 6.0% and that for the covariate adjusted models had a bias of 1.2% (Table 16). In the RIE attrition dataset, at 3-Day follow-up time period, the unadjusted models had a bias 15.7%. Bias was reduced to 11.7% and 14.9% when models were interaction adjusted and covariate adjusted, respectively. The bias reduction was moderate in the interaction adjusted models (Table 17). There was minimal bias reduction in the RIET attrition dataset. There was a 19.0% bias in the mean treatment effect from the unadjusted models and a 20.1% and a 16.5% bias in the interaction adjusted and covariate adjusted models respectively (Table 18).

Table 16: Approach 1 – Impact of Propensity Scores using Sum of Observations in the RI Dataset

Description of Parameter Estimators	Complete Dataset 5000 Simulations	Coefficient Means Unadjusted Models	Coefficient Means Interaction Adjusted	Coefficient Means Covariate Adjusted
Intercept	5.3803	5.5691	5.5582	5.5565
Group (Experimental Group)	-0.0348	-0.0514	-0.0470	-0.0473
Time Period (3-Day)	0.0625	0.0508	0.0481	0.0517
Time Period (3-Month)	0.1031	0.1345	0.1241	0.1354
Propensity Score	*	*	1.6646	-0.9768
Gender (Female)	0.1460	0.0451	0.0485	0.0492
Treatment Effect (3-Day and Exp. Group)	0.1714	0.1749	0.1817	0.1734
Treatment Effect (3-Month and Exp. Group)	0.2047	0.1973	0.2073	0.1963

Table 17: Approach 1 – Impact of Propensity Scores using Sum of Observations in the RIE Dataset

Description of Parameter Estimators	Complete Dataset 5000 Simulations	Coefficient Means Unadjusted Models	Coefficient Means Interaction Adjusted	Coefficient Means Covariate Adjusted
Group (Experimental Group)	-0.0348	-0.1000	-0.0440	-0.0458
Time Period (3-Day)	0.0625	0.0449	0.0453	0.0449
Time Period (3-Month)	0.1031	0.1344	0.1294	0.1353
Propensity Score	*	*	-5.6185	-5.9252
Gender (Female)	0.1460	0.0701	0.0810	0.0817
Treatment Effect (3-Day and Exp. Group)	0.1714	0.1983	0.1915	0.1969
Treatment Effect (3-Month and Exp. Group)	0.2047	0.2090	0.2112	0.2080

Table 18: Approach 1 – Impact of Propensity Scores using Sum of Observations in the RIET Dataset

Description of Parameter Estimators	Complete Dataset 5000 Simulations	Coefficient Means Unadjusted Models	Coefficient Means Interaction Adjusted	Coefficient Means Covariate Adjusted
Intercept	5.3803	5.7072	5.8626	5.8596
Group (Experimental Group)	-0.0348	-0.1205	-0.1213	-0.1127
Time Period (3-Day)	0.0625	0.0409	0.0441	0.0416
Time Period (3-Month)	0.1031	0.1284	0.1281	0.1294
Propensity Score	*	*	-7.5290	-7.1348
Gender (Female)	0.1460	0.0618	0.0727	0.0725
Treatment Effect (3-Day and Exp. Group)	0.1714	0.2040	0.2059	0.1997
Treatment Effect (3-Month and Exp. Group)	0.2047	0.1712	0.1872	0.1848

Propensity Scores Computed using Missing Data Pattern at 3-Day

Similar to propensity scores estimated from random intercept and from the sum of observations using the RI attrition dataset, propensity scores estimated from missing data patterns using RI attrition dataset had minimal bias. The mean treatment effect bias at 3-Day for the unadjusted models was 2%, while the bias for the interaction adjusted model and covariate adjusted models were 6.9% and 1.5%. All models had a bias that fell within the $\pm 10\%$ threshold (Table 19).

On the other hand, for the RIE attrition dataset, the treatment effect for the unadjusted models had a bias of 15.7%. The interaction adjusted models reduced the bias to 12.3% and covariate adjusted models reduced the bias to 14.5% (Table 20). In the RIET attrition dataset, the mean treatment effects for unadjusted models had a bias of 19.0%. The interaction adjusted models reduced the bias to 17.8% and covariate adjusted models reduced the bias to 18.7%. The weak response of bias in this case be attributed to the short follow-up period (baseline, 3-Day and 3-Month), hence making the longitudinal component rather short (Table 21).

Table 19: Approach 1 – Impact of Propensity Scores using Missing Data Patterns in the RI Dataset

Description of Parameter Estimators	Complete Dataset 5000 Simulations	Coefficient Means Unadjusted Models	Coefficient Means Interaction Adjusted	Coefficient Means Covariate Adjusted
Intercept	5.3803	5.5691	5.5594	5.5612
Group (Experimental Group)	-0.0348	-0.0514	-0.0434	-0.0467
Time Period (3-Day)	0.0625	0.0508	0.0558	0.0508
Time Period (3-Month)	0.1031	0.1345	0.1374	0.1372
Propensity Score	*	*	-0.6950	-0.4232
Gender (Female)	0.1460	0.0451	0.0461	0.0467
Treatment Effect (3-Day and Exp. Group)	0.1714	0.1749	0.1833	0.1740
Treatment Effect (3-Month and Exp. Group)	0.2047	0.1973	0.2040	0.1975

Table 20: Approach 1 – Impact of Propensity Scores using Missing Data Patterns in the RIE Dataset

Description of Parameter Estimators	Complete Dataset 5000 Simulations	Coefficient Means Unadjusted Models	Coefficient Means Interaction Adjusted	Coefficient Means Covariate Adjusted
Intercept	5.3803	5.7067	5.6759	5.6769
Group (Experimental Group)	-0.0348	-0.1000	-0.0468	-0.0451
Time Period (3-Day)	0.0625	0.0449	0.0572	0.0469
Time Period (3-Month)	0.1031	0.1344	0.1328	0.1348
Propensity Score	*	*	-3.6617	-3.6037
Gender (Female)	0.1460	0.0701	0.0797	0.0801
Treatment Effect (3-Day and Exp. Group)	0.1714	0.1983	0.1926	0.1962
Treatment Effect (3-Month and Exp. Group)	0.2047	0.2090	0.2225	0.2083

Table 21: Approach 1 – Impact of Propensity Scores using Missing Data Patterns in the RIET Dataset

Description of Parameter Estimators	Complete Dataset 5000 Simulations	Coefficient Means Unadjusted Models	Coefficient Means Interaction Adjusted	Coefficient Means Covariate Adjusted
Intercept	5.3803	5.7072	5.7354	5.7041
Group (Experimental Group)	-0.0348	-0.1205	-0.1170	-0.1095
Time Period (3-Day)	0.0625	0.0409	0.0763	0.0394
Time Period (3-Month)	0.1031	0.1284	0.2357	0.1374
Propensity Score	*	*	0.5442	-0.2803
Gender (Female)	0.1460	0.0618	0.0681	0.0621
Treatment Effect (3-Day and Exp. Group)	0.1714	0.2040	0.1980	0.2036
Treatment Effect (3-Month and Exp. Group)	0.2047	0.1712	0.2181	0.1716

The performance of propensity scores at 3-Day follow-up and under Approach 1, estimated using random intercept, sum of observations, and missing data patterns across the three types of attrition datasets (RI, RIE, and RIET) is summarized below (Figure 22). The results included comparisons across different types of models (unadjusted, covariate adjusted, and interaction adjusted). Propensity scores estimated using random intercept are effective in reducing treatment effect bias regardless of the attrition dataset. However, the effectiveness of reducing bias is limited to the interaction adjusted models. The covariate adjusted models minimally affected the mean treatment effect bias from unadjusted models regardless of how the propensity scores were computed (random intercept, number of observations, and missing data patterns). Propensity scores estimated from the sum of observations and from missing data patterns were more effective in reducing bias in the RIE attrition dataset than they were in the RIET attrition dataset.

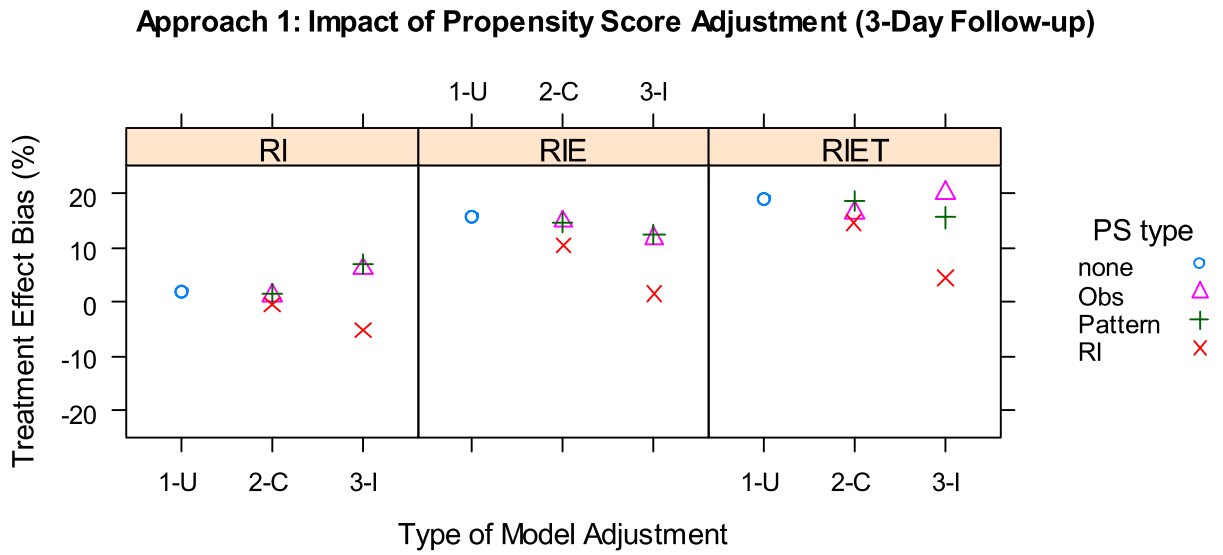

Figure 22: Propensity scores estimated using random intercept, number of observations, and missing data patterns are applied in the three attrition datasets (RI, RIE, and RIET).

Propensity Scores Computed using Random Intercept at 3-Months

The propensity scores estimated using random intercept were effective in either keeping bias or reducing bias to within the ±10% threshold across all attrition datasets. For the RI attrition dataset, the unadjusted models underestimated the mean treatment effect by -3.6%, while the covariate adjusted and interaction adjusted models underestimated the mean treatment effect by -6.9% and -10% respectively (Table 13). In the RIE attrition dataset, the mean treatment effect from the unadjusted models had a 2.1% bias, which was reduced to 0.7% by the covariate adjusted models, and increased to -7.7% for the interaction adjusted models. However, the bias across all models remained within a ±10% threshold (Table 14). There was a -16.4% bias in the mean treatment effect for the unadjusted models using the RIET attrition dataset. This interaction adjusted models reduced the bias from -16.4% to -2.9%, while the covariate adjusted model reduced the bias from -16.4% to 3.8% (Table 15).

Propensity Scores Computed using the Number of Observations at 3-Months

In general, the propensity scores estimated using the numbers of observations were effective at reducing bias. In the RI dataset, the interaction adjusted models reduced bias from -3.6% to 1.3% and the covariate adjusted models underestimated the mean treatment effect by 4.1% compared to 3.6% underestimation of the mean treatment effect in the unadjusted models (Table 16). The mean treatment effect in the unadjusted models using the RIE dataset had a bias of 2.1% while the mean treatment effect in the interaction adjusted models and the covariate adjusted models had a bias of 3.2% and 1.6% respectively (Table 17). In the RIET dataset, the mean treatment effect in the unadjusted models had a bias of -16.4%. This bias was reduced to -8.6% in the interaction adjusted models and to -9.7% in the covariate adjusted models (Table 18).

Propensity Scores Computed using the Missing Data Patterns at 3-Months

In the case of propensity scores estimated from missing data patterns using the RI attrition dataset, mean treatment effect from unadjusted models had a bias of -3.6% while the interaction adjusted models had a bias of -0.3% and covariate adjusted models had a bias of -3.5% (Table 19). In the RIE attrition dataset, the bias from the unadjusted models was 2.1% while the bias from the interaction adjusted models was 8.7%, which was within the $\pm 10\%$ threshold. The covariate adjusted models had a bias of 1.8% (Table 20). Lastly, in the RIET attrition dataset, the mean treatment effect estimated from the unadjusted models and from the covariate adjusted models were essentially equivalent at -16.4% and -16.2%, respectively. However, interaction adjusted models reduced the bias from -16.4% to 6.6% (Table 21).

The performance of the propensity scores under Approach 1, for the 3-Month follow-up period, estimated using either random intercept, number of observations or pattern mixture is summarized below (Figure 23). Across all three types of attrition datasets and for both covariate adjusted and interaction adjusted models, propensity scores using RI performed the best in bias reduction. Unlike the 3-Day follow-up period, the propensity scores using both number of observations and the missing data patterns either keep bias within the $\pm 10\%$ threshold or effectively reduced bias in the attrition datasets. Of the three types of propensity scores, those estimated from missing data patterns were least effective in reducing bias. It is unclear whether this would be the same if the longitudinal component of the CRT were longer.

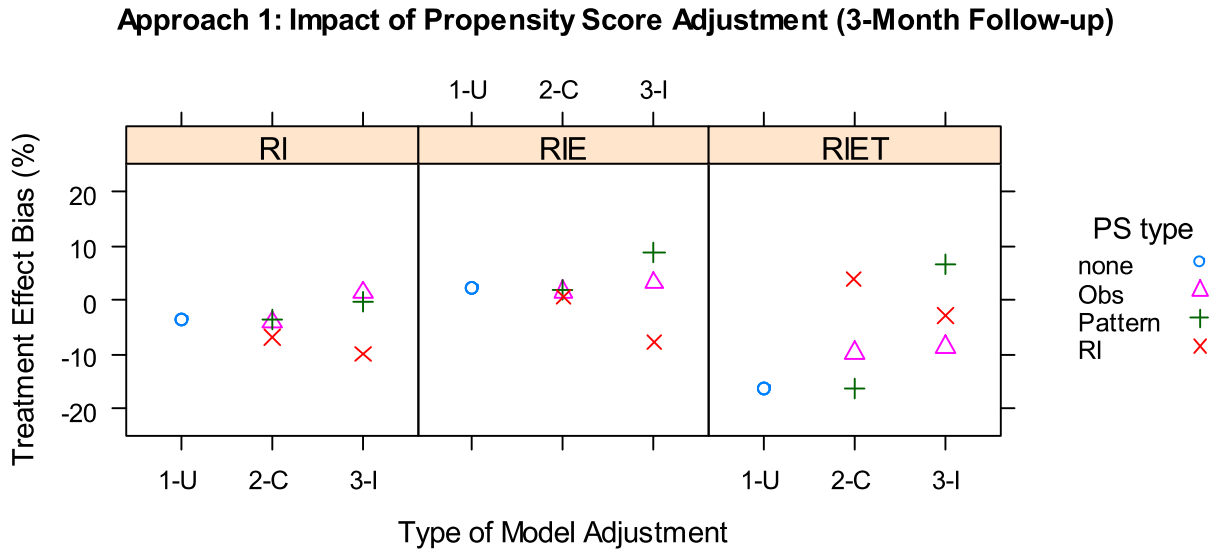

Figure 23: Propensity scores estimated using random intercept, number of observations, and missing data patterns are applied in the three attrition datasets (RI, RIE, and RIET).

4.2.2 Performance of Propensity Scores under Approach 2

Propensity Scores Computed using Random Intercept at 3-Day

The performances of propensity scores estimated under Approach 2, in which individuals with missing outcome data in all three periods are excluded, were investigated (Figure 8). As was the case in Approach 1, the bias in the RI attrition dataset under Approach 2 was minimal. The mean treatment effect from the unadjusted models had a bias of 2.0%. The covariate adjusted models reduced the bias from 2.0% to -0.4%. The mean treatment effects of the interaction adjusted models had a bias of-3.2% (Table 22). In the RIE attrition dataset, bias in mean treatment effect estimates was 15.7%. The interaction adjusted models reduced the bias to 6.2%, while the covariate adjusted model reduced the bias to 10.5% (Table 23). The mean treatment effect for the unadjusted models in the RIET attrition dataset had a bias of 19.0%. This bias was reduced to 7.7%, within the $\pm 10\%$ threshold, by the interaction adjusted models and reduced to 14.5% by the covariate adjusted models (Table 24).

Table 22: Approach 2 – Impact of Propensity Scores using Random Intercept in the RI Dataset

Description of Parameter Estimators	Complete Dataset 5000 Simulations	Coefficient Means Unadjusted Models	Coefficient Means Interaction Adjusted	Coefficient Means Covariate Adjusted
Intercept	5.3803	5.5691	5.5067	5.5080
Group (Experimental Group)	-0.0348	-0.0514	0.0035	0.0018
Time Period (3-Day)	0.0625	0.0508	0.0566	0.0521
Time Period (3-Month)	0.1031	0.1345	0.1400	0.1369
Propensity Score	*	*	-32.0540	-31.9709
Gender (Female)	0.1460	0.0451	0.1333	0.1333
Treatment Effect (3-Day and Exp. Group)	0.1714	0.1749	0.1659	0.1708
Treatment Effect (3-Month and Exp. Group)	0.2047	0.1973	0.1864	0.1905

Table 23: Approach 2 – Impact of Propensity Scores using Random Intercept in the RIE Dataset

Description of Parameter Estimators	Complete Dataset 5000 Simulations	Coefficient Means Unadjusted Models	Coefficient Means Interaction Adjusted	Coefficient Means Covariate Adjusted
Intercept	5.3803	5.7067	5.6158	5.6142
Group (Experimental Group)	-0.0348	-0.1000	-0.0098	-0.0075
Time Period (3-Day)	0.0625	0.0449	0.0568	0.0506
Time Period (3-Month)	0.1031	0.1344	0.1388	0.1315
Propensity Score	*	*	-15.3474	-15.6269
Gender (Female)	0.1460	0.0701	0.1395	0.1388
Treatment Effect (3-Day and Exp. Group)	0.1714	0.1983	0.1821	0.1895
Treatment Effect (3-Month and Exp. Group)	0.2047	0.2090	0.1974	0.2061

Table 24: Approach 2 – Impact of Propensity Scores using Random Intercept in the RIET Dataset

Description of Parameter Estimators	Complete Dataset 5000 Simulations	Coefficient Means Unadjusted Models	Coefficient Means Interaction Adjusted	Coefficient Means Covariate Adjusted
Intercept	5.3803	5.7072	5.5925	5.5907
Group (Experimental Group)	-0.0348	-0.1205	-0.0132	-0.0099
Time Period (3-Day)	0.0625	0.0409	0.0538	0.0423
Time Period (3-Month)	0.1031	0.1284	0.1335	0.1240
Propensity Score	*	*	-10.0602	-10.2865
Gender (Female)	0.1460	0.0618	0.1388	0.1387
Treatment Effect (3-Day and Exp. Group)	0.1714	0.2040	0.1846	0.1962
Treatment Effect (3-Month and Exp. Group)	0.2047	0.1712	0.2056	0.2121

Propensity Scores Computed using Sum of Observations at 3-Day

The performance of propensity scores estimated using sum of observations under Approach 2 will be examined in this section. The unadjusted mean treatment effect estimate for the RI attrition dataset indicates a bias of 2.0%. The mean treatment effect for the interaction adjusted models and covariate adjusted models had a bias of 8.1% and 1.2% respectively. In this case, all bias remained within the $\pm 10\%$ threshold (Table

25). However, in the RIE attrition dataset, the mean treatment effect estimate for the unadjusted models and the covariate adjusted models were marginally different, 15.7% and 14.9% respectively. The bias was reduced from 15.7% to 12.9% in the interaction adjusted models (Table 26). In the RIET attrition dataset, the estimated treatment effect had a bias of 19.0%, which was reduced to 8.1% by the applying the propensity scores as an interaction term and to 16.5% by applying the propensity scores as a covariate (Table 27).

Table 25: Approach 2 - Impact of Propensity Scores using Sum of Observations in the RI Dataset

Description of Parameter Estimators	Complete Dataset 5000 Simulations	Coefficient Means Unadjusted Models	Coefficient Means Interaction Adjusted	Coefficient Means Covariate Adjusted
Intercept	5.3803	5.7072	5.5925	5.5907
Group (Experimental Group)	-0.0348	-0.1205	-0.0132	-0.0099
Time Period (3-Day)	0.0625	0.0409	0.0538	0.0423
Time Period (3-Month)	0.1031	0.1284	0.1335	0.1240
Propensity Score	*	*	-10.0602	-10.2865
Gender (Female)	0.1460	0.0618	0.1388	0.1387
Treatment Effect (3-Day and Exp. Group)	0.1714	0.2040	0.1846	0.1962
Treatment Effect (3-Month and Exp. Group)	0.2047	0.1712	0.2056	0.2121

Table 26: Approach 2 - Impact of Propensity Scores using Sum of Observations in the RIE Dataset

Description of Parameter Estimators	Complete Dataset 5000 Simulations	Coefficient Means Unadjusted Models	Coefficient Means Interaction Adjusted	Coefficient Means Covariate Adjusted
Group (Experimental Group)	-0.0348	-0.0405	-0.0535	-0.0473
Time Period (3-Day)	0.0625	0.0401	0.0489	0.0517
Time Period (3-Month)	0.1031	0.1330	0.1302	0.1354
Propensity Score	*	*	1.6646	-0.9768
Gender (Female)	0.1460	0.0451	0.0485	0.0492
Treatment Effect (3-Day and Exp. Group)	0.1714	0.1749	0.1854	0.1734
Treatment Effect (3-Month and Exp. Group)	0.2047	0.1973	0.2033	0.1963

Table 27: Approach 2 - Impact of Propensity Scores using Sum of Observations in the RIET Dataset

Description of Parameter Estimators	Complete Dataset 5000 Simulations	Coefficient Means Unadjusted Models	Coefficient Means Interaction Adjusted	Coefficient Means Covariate Adjusted
Intercept	5.3803	5.7067	5.6603	5.6524
Group (Experimental Group)	-0.0348	-0.1000	-0.0488	-0.0458
Time Period (3-Day)	0.0625	0.0449	0.0405	0.0449
Time Period (3-Month)	0.1031	0.1344	0.1256	0.1353
Propensity Score	*	*	-5.6185	-5.9252
Gender (Female)	0.1460	0.0701	0.0810	0.0817
Treatment Effect (3-Day and Exp. Group)	0.1714	0.1983	0.1935	0.1969
Treatment Effect (3-Month and Exp. Group)	0.2047	0.2090	0.2133	0.2080

Propensity Scores Computed using Missing Data Patterns at 3-Day

The propensity scores in this case were effective at reducing bias when using the RI attrition dataset but not in the RIE or RIET attrition dataset. In the RI dataset, the bias in the mean treated effect of the unadjusted models was 2.0%, while the bias in the interaction and covariate adjusted models was 4.0% and 1.5% respectively (Table 28). The mean treatment effect in the unadjusted models for the RIE dataset had a bias of 15.7% which was reduced to 12.4% and to 14.5% by adding the propensity score as an interaction term and covariate term respectively (Table 29). In the RIET dataset, the bias in the mean treatment effect for the unadjusted, interaction adjusted and covariate adjusted models was 19.0%, 17.8% and 18.7% (Table 30).

Table 28: Approach 2 - Impact of Propensity Scores using Missing Data Pattern in the RI Dataset

Description of Parameter Estimators	Complete Dataset 5000 Simulations	Coefficient Means Unadjusted Models	Coefficient Means Interaction Adjusted	Coefficient Means Covariate Adjusted
Intercept	5.3803	5.7072	5.6753	5.6832
Group (Experimental Group)	-0.0348	-0.1205	-0.1009	-0.1127
Time Period (3-Day)	0.0625	0.0409	0.0526	0.0416
Time Period (3-Month)	0.1031	0.1284	0.1426	0.1294
Propensity Score	*	*	-7.5290	-7.1348
Gender (Female)	0.1460	0.0618	0.0727	0.0725
Treatment Effect (3-Day and Exp. Group)	0.1714	0.2040	0.1854	0.1997
Treatment Effect (3-Month and Exp. Group)	0.2047	0.1712	0.1660	0.1848

Table 29: Approach 2 - Impact of Propensity Scores using Missing Data Pattern in the RIE Dataset

Description of Parameter Estimators	Complete Dataset 5000 Simulations	Coefficient Means Unadjusted Models	Coefficient Means Interaction Adjusted	Coefficient Means Covariate Adjusted
Intercept	5.3803	5.7067	5.6607	5.6602
Group (Experimental Group)	-0.0348	-0.1000	-0.0485	-0.0451
Time Period (3-Day)	0.0625	0.0449	0.0573	0.0469
Time Period (3-Month)	0.1031	0.1344	0.1288	0.1348
Propensity Score	*	*	-3.6617	-3.6037
Gender (Female)	0.1460	0.0701	0.0797	0.0801
Treatment Effect (3-Day and Exp. Group)	0.1714	0.1983	0.1927	0.1962
Treatment Effect (3-Month and Exp. Group)	0.2047	0.2090	0.2239	0.2083

Table 30: Approach 2 - Impact of Propensity Scores using Missing Data Pattern in the RIET Dataset

Description of Parameter Estimators	Complete Dataset 5000 Simulations	Coefficient Means Unadjusted Models	Coefficient Means Interaction Adjusted	Coefficient Means Covariate Adjusted
Intercept	5.3803	5.7072	5.7479	5.6973
Group (Experimental Group)	-0.0348	-0.1205	-0.1308	-0.1095
Time Period (3-Day)	0.0625	0.0409	0.0908	0.0394
Time Period (3-Month)	0.1031	0.1284	0.1822	0.1374
Propensity Score	*	*	0.5442	-0.2803
Gender (Female)	0.1460	0.0618	0.0681	0.0621
Treatment Effect (3-Day and Exp. Group)	0.1714	0.2040	0.2019	0.2036
Treatment Effect (3-Month and Exp. Group)	0.2047	0.1712	0.2213	0.1716

The performance of propensity scores estimated using either random intercept, number of observations, or missing data patterns for the 3-Day follow-up period under approach 2 is summarized below (Figure 24). Here we find that the covariate adjusted models have the same level of bias as the unadjusted models across the three different types of attrition datasets regardless of the type of propensity score used. Also, the bias in the RI attrition dataset was negligible and propensity score adjustments kept the bias within the ±10% threshold. This characteristic was also found in Approach 1 in both the 3-Day and 3-Month follow-up results. Applying propensity scores as an interaction term in the models was effective in reducing the mean treatment effect bias from the unadjusted models, especially when propensity scores were estimated using random intercept or the sum of observations.

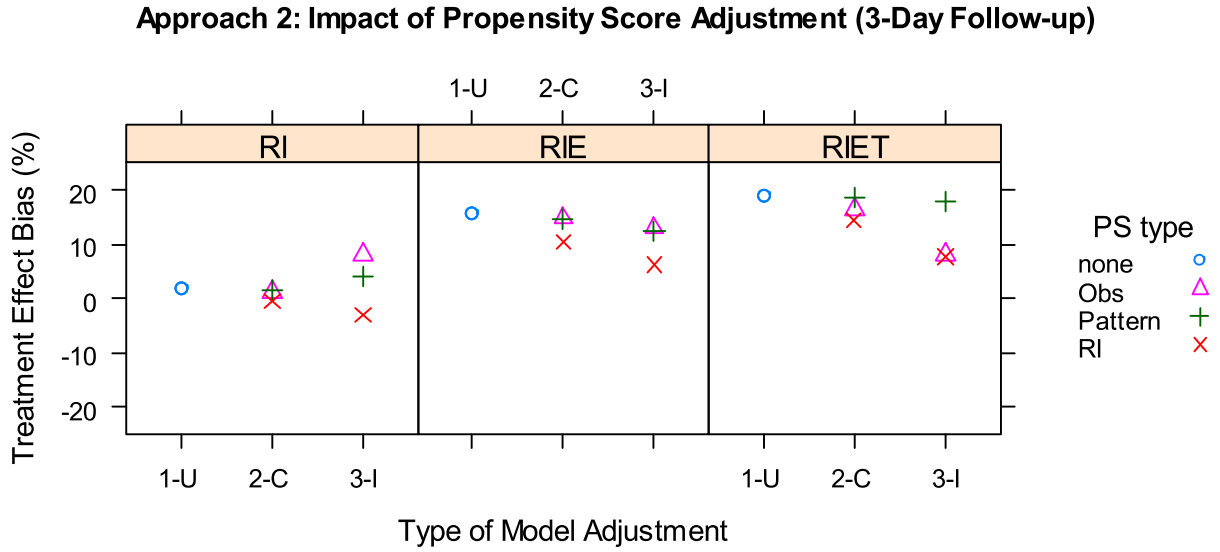

Figure 24: Propensity scores estimated using random intercept, number of observations, and missing data patterns are applied in the three attrition datasets (RI, RIE, and RIET).

Propensity Scores Computed using Random Intercept at 3-Month

In the RI attrition dataset, the unadjusted mean treatment effect had a bias -3.6%, while the interaction adjusted models and covariate adjusted models had bias of -9.0% and -6.9% (Table 22). The bias across the different types of model consistently stayed within the $\pm 10\%$ threshold as in the all other results reported. Also, the bias across the different types of models in the RIE dataset remained very low. The mean treatment effect in the unadjusted models and the covariate adjusted models were overestimated by 2.1% and 0.7% respectively. However, the mean treatment effect in the interaction models was underestimated by only 3.5% (Table 23). In the RIET attrition dataset, the mean treatment effect in the unadjusted model had a bias of -16.4%. However, the bias was effectively reduced by the interaction adjusted and covariate adjusted models to 0.4% and to 3.6%, respectively (Table 24).

Propensity Scores Computed using Sum of Observations at 3-Month

A low level of bias was found in the RI and RIE attrition dataset. In the RI dataset, the average treatment effect of the unadjusted models had a bias of -3.6%. This bias was reduced to -0.7% when the propensity scores estimated using the sums of observations were applied to the models as an interaction. When the propensity scores were applied as a covariate the bias was still low at -4.1% (Table 25). In the RIE attrition dataset, the average treatment effect, estimated using unadjusted models, had a bias of 2.1%. This was reduced to 1.6% in the covariate adjusted models. In the interaction adjusted models the bias remained low at 4.2% (Table 26). The bias in the RIET attrition dataset was high. The average treatment effect for the unadjusted models had a bias of -16.4%.

The bias was reduced in the covariate adjusted models from -16.4% to -9.7%. However, in the interaction adjusted models, the bias increased marginally from -16.4% to -18.9% (Table 27).

Propensity Scores Computed using Missing Data Pattern at 3-Month

Lastly, the performance of propensity scores estimated using missing data patterns is investigated. In the RI attrition dataset, the mean treatment effect estimate for the unadjusted models had a bias of -3.6%. Bias in the models that applied propensity scores as an interaction and that applied propensity scores as a covariate reduced bias to -1.0% and -3.5% respectively (Table 28). In the RIE dataset, the bias in the unadjusted models and the covariate adjusted models was 2.1% and 1.8%. The bias in the models that applied the propensity scores as an interaction was 9.4% (Table 29). In the RIET attrition dataset, the bias in the unadjusted models was high at -16.4%. This bias was reduced to 8.1% when the models applied propensity scores as an interaction. However, applying the propensity scores in model as a covariate had very little impact on the bias. The bias remained at -16.2% (Table 30).

The performance of propensity scores for 3-Month follow-up under Approach 2 estimated using either random intercept, number of observations or missing data patterns is summarized below (Figure 25). The bias across all types of models in the RI attrition dataset remained within the acceptable bias threshold of $\pm 10\%$. The propensity scores estimated using a random intercept were the most effective at reducing bias followed by propensity estimated using the sum of observations. The effectiveness of

bias reduction varied the most in RIET attrition dataset, which also had the highest level of bias in the mean treatment effect of the unadjusted models.

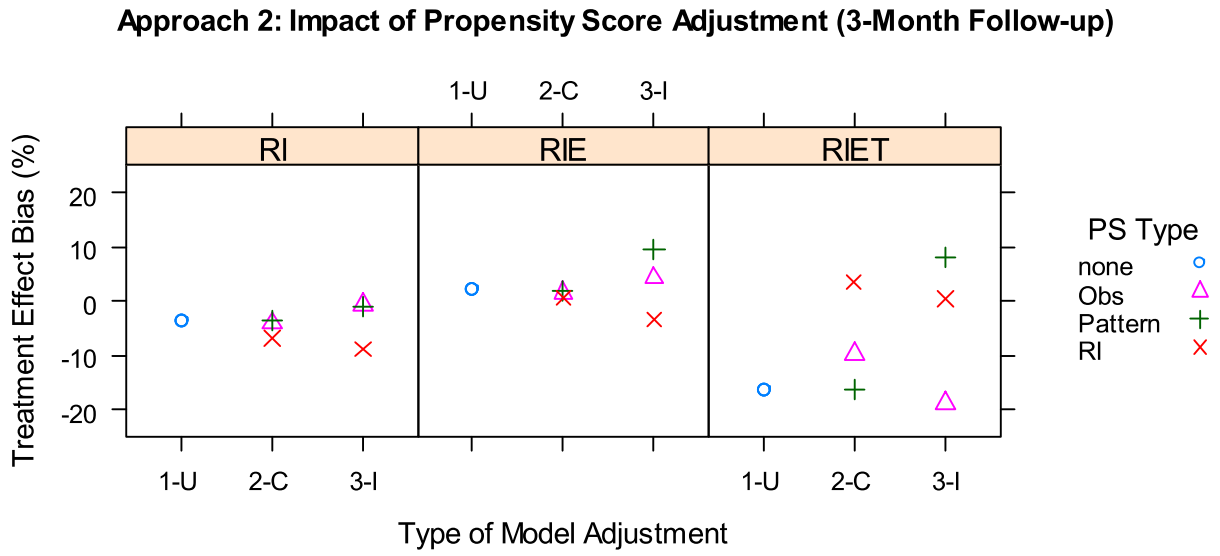

Figure 25: Propensity scores estimated using random intercept, number of observations, and missing data patterns are applied in the three attrition datasets (RI, RIE, and RIET).

4.3 Case Study Results

The developed methods were applied to the case study. The bias literacy study was characterized with low level of bias. This was partly attributed to a well-designed pair matching strategy that balanced baseline covariates. The bias literacy data allowed for the application of only Approach 2 because data were only collected on respondents (Figure 8). The results with the exception of one instance are consistent with those described earlier.

Minimal bias is found in the treatment effect estimates of the bias literacy CRT when compared to mean treatment effect of 5000 simulations of the complete dataset. In this case, the bias literacy CRT will be referred to as the unadjusted model. The bias literacy CRT before adjustment had a bias of -3.2% for the 3-Day treatment effect and 0.4% for

the 3-Month treatment effect. When propensity scores estimated from random effects were applied as an interaction, the bias remained within the ±10% threshold at -1.7% for the 3-Day follow-up time period and 8.2% for the 3-Month follow-up time period (Table 31). On the other hand, the bias was -3.5% for the 3-Day time period and 1.1% for the 3-Month time period when the propensity scores were applied to the model as a covariate.

In the case when propensity scores were estimated using a missing data pattern, the 3-Day follow-up bias was -5.6%. However, in the 3-Month period, the bias was -38.0%. This increase in bias may be attributed to the very short longitudinal component. The covariate adjusted model had a bias of -3.5% and 1.6% for the 3-Day time period and 3-Month time period, respectively (Table 32). In the case when propensity scores were estimated using the sum of observations method and applied as an interaction term, the bias in the 3-Day follow-up period was -4.9% and in the 3-Month follow-up period was 2.5% (Table 33). When the propensity score was applied as a covariate term, the bias was -4.5% in the 3-Day follow-up period and -1.4% in the 3-Month follow-up period. In general, the results were very consistent with those discussed earlier. Propensity scores estimated using random effects were most effective in reducing bias, followed by those estimated by the number of observations. Propensity scores estimated using the missing data patterns had a few mixed results. This calls for further research using a longer longitudinal component for CRT.

Table 31: Case Study Results - Impact of Propensity Scores using Missing Data Pattern

Description of Parameter Estimators	Complete Dataset 5000 Simulations	Bias Literacy CRT Dataset	Interaction Adjusted Model	Covariate Adjusted Model
Intercept	5.3803	5.3912	5.3859	5.3837
Group (Experimental Group)	-0.0348	-0.0391	-0.0425	-0.0393
Time Period (3-Day)	0.0625	0.0673	0.0648	0.0668
Time Period (3-Month)	0.1031	0.1072	0.1015	0.1055
Propensity Score	*	*	6.8032	3.8499
Gender (Female)	0.1460	0.1543	0.1463	0.1473
Treatment Effect (3-Day and Exp. Group)	0.1714	0.1659	0.1685	0.1654
Treatment Effect (3-Month and Exp. Group)	0.2047	0.2055	0.2215	0.2070

Table 32: Case Study Results - Impact of Propensity Scores using Missing Data Patterns

Description of Parameter Estimators	Complete Dataset 5000 Simulations	Bias Literacy CRT Dataset	Interaction Adjusted Model	Covariate Adjusted Model
Intercept	5.3803	5.3912	5.3997	5.3832
Group (Experimental Group)	-0.0348	-0.0391	-0.0137	-0.0373
Time Period (3-Day)	0.0625	0.0673	0.0457	0.0642
Time Period (3-Month)	0.1031	0.1072	0.1205	0.1025
Propensity Score	*	*	-0.3041	-0.0479
Gender (Female)	0.1460	0.1543	0.1487	0.1497
Treatment Effect (3-Day and Exp. Group)	0.1714	0.1659	0.1618	0.1655
Treatment Effect (3-Month and Exp. Group)	0.2047	0.2055	0.1269	0.2080

Table 33: Case Study Results - Impact of Propensity Scores using Sum of Observations

Description of Parameter Estimators	Complete Dataset 5000 Simulations	Bias Literacy CRT Dataset	Interaction Adjusted Model	Covariate Adjusted Model
Intercept	5.3803	5.3912	5.3680	5.3701
Group (Experimental Group)	-0.0348	-0.0391	-0.0243	-0.0257
Time Period (3-Day)	0.0625	0.0673	0.0728	0.0715
Time Period (3-Month)	0.1031	0.1072	0.1178	0.1088
Propensity Score	*	*	-2.7969	-3.1778
Gender (Female)	0.1460	0.1543	0.1590	0.1591
Treatment Effect (3-Day and Exp. Group)	0.1714	0.1659	0.1631	0.1638
Treatment Effect (3-Month and Exp. Group)	0.2047	0.2055	0.1995	0.2019

CHAPTER 5
DISCUSSION

5.1 Discussion of Findings

This study demonstrated that different types of non-ignorable missingness have different effects on bias in a cluster randomized trial. Missingness that depends only on random effects, while non-ignorable, induces negligible levels of bias because the missingness impacts the different groups similarly. Bias is amplified when missingness that depends on random effects differentially affects groups. This type of non-ignorable missing data mechanism exists when the magnitude of missingness differs based on distinct group characteristics. In this study, the characteristics were either the experimental group or the experimental group and follow-up time.

This research confirms that missingness depending on random effects can be captured by missingness patterns and the number of observations (Park, et al., 2002). This was illustrated by the linear relationship between the propensity scores estimated using random effect and propensity scores estimated using the number of observations or estimated using missing data patterns. Furthermore, propensity scores under Approach 1 and those under Approach 2 were also linearly related to each other. Hence, both approaches were expected to have similar outcomes in bias reduction. However, compared to the propensity scores estimated using random effects, those estimated using the number of observations and missing data patterns showed more variation.

This is because there were only three observations for each individual. The variation may also be attributed to the differential attrition in the RIE and RIET datasets. Propensity scores that, in addition to missing data characteristics, incorporated cluster level and individual variables were not effective in adjusting bias. Consequently, only propensity scores estimated from missing data characteristics were applied in this study.

The research findings suggest that applying propensity scores as an interaction term as opposed to a covariate term is more effective in reducing bias. This is because non-ignorable missingness induces bias only when missingness differentially affects groups, which can be described as a form of interaction. Propensity scores adjusted as covariates did not significantly change the bias found in unadjusted models. The study also established that propensity scores under Approach 1 and under Approach 2 had similar levels of performance. This result suggests that propensity scores aimed at correcting bias from non-ignorable missingness need not be informed by demographic information or other variables but rather should focus on the missing data characteristics. The results also show that propensity scores estimated using random effects are most effective at reducing bias. However, in practice, random effects cannot be assumed to be correctly estimable. The results indicate that propensity scores estimated from number of observations and missing data patterns, while not as effective at reducing bias as propensity scores estimated using random effects, were moderately successful at bias reduction. The moderate success may be attributed to the very short longitudinal component (3 observations for each individual) of the cluster randomized trial.

In observational studies, propensity scores have not only been used as an adjustment method for selection bias reduction (Paul R. Rosenbaum & Rubin, 1983) but have also been used to give observational studies characteristics of randomized trials (D'Agostino, 1998; Williamson, Morley, Lucas, & Carpenter, 2012). CRT's, despite being randomized trials, remain vulnerable to bias because the randomization is at a cluster level as opposed to an individual (unit) level. Opportunity for differential attrition in CRTs with longitudinal data also further contributes to bias. The incorporation of the number of observations and missing data patterns into propensity scores, which were subsequently applied to the analysis utilizing mixed effects models, is a novel idea in this research. The study establishes that propensity score based methods can be used to address bias from differential attrition in CRT's with a longitudinal component. Cluster randomized trials are gaining more acceptance as an appropriate industrial engineering tool for studying interventions aimed at improving quality of healthcare processes. This research contributes pragmatic ways of addressing bias in such trials.

5.2 Study Limitations

The research study did not use a cluster randomized trial data with a sufficiently long longitudinal component to allow a deeper investigation into the full potential of propensity scores estimated from the sum of observations or missing data patterns in correcting bias. Also, the amplification of random effects was not based on established research information published on intraclass correlation. Furthermore, an assumption was made about the degree of similarity between bias literacy cluster randomized trial in an academic environment and a cluster randomized trial in the context of a healthcare

system. By making this assumption, it is implied that the methods investigated in this research can be seamlessly applied to improving the quality of healthcare system processes. In practice, the two types of environments may have different characteristics. On the other hand, differential attrition is a common feature of any experimental design with longitudinal data. Hence, bias reduction methods developed in this work may be relevant to other experimental designs utilizing longitudinal data. Also, a limited set of parameters were used for the simulations. While the parameters were applied in a CRT setting, more complex cluster level structure may need to be addressed.

5.3 Future Research

In future research, the use of weights, at the unit and cluster level, appropriate for multilevel models needs to be investigated for their effectiveness of reduction of bias in cluster randomized trials with a longitudinal component. This would incorporate more complex cluster level structure that may arise in practice. The investigation would include a survey and comparison of current software in their abilities to accommodate such weights. Furthermore, it would be beneficial to investigate the effectiveness of using a combination of other methods for controlling for baseline imbalance by use of stratification, while separately addressing attrition using propensity based methods.

REFERENCES:

Abel, U., & Koch, A. (1999). The role of randomization in clinical studies: myths and beliefs. *J Clin Epidemiol, 52*(6), 487-497.

Aguayo, R. (1991). *Dr. Deming: The American who taught the Japanese about quality*: Simon and Schuster.

Airy, G. B. (1861). *On the algebraical and numerical theory of errors of observations and the combination of observations*: Macmillan and Company.

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

Anderson, R. (2008). New MRC guidance on evaluating complex interventions. *BMJ, 337*. doi: 10.1136/bmj.a1937

Attewell, A. (1998). Florence Nightingale (1820-1910). *Prospects, 28*(1), 151-166. doi: http://dx.doi.org/10.1007/BF02737786

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review, 84*(2), 191-215. doi: 10.1037/0033-295x.84.2.191

Bandura, A. (1991). Social cognitive theory of self-regulation. *Organizational Behavior and Human Decision Processes, 50*(2), 248-287. doi: http://dx.doi.org/10.1016/0749-5978(91)90022-L

Bandura, A. (1997). Self-efficacy: The exercise of control: New York: Freeman.

Bandura, A., & Kazdin, A. E. (2000). Social-cognitive theory *Encyclopedia of psychology, Vol. 7.* (pp. 329-332). Washington, DC, New York, NY, USUS: American Psychological Association

Oxford University Press.

Barness, Z., Shortell, S., Gillies, R., Hughes, E., O'Brien, J., Bohr, D., . . . Kralovec, P. (1993). The quality march. *Hospitals and Health Networks, 5*, 52-55.

Baron, J. H. (2012). Evolution of clinical research: A history before and beyond James Lind. *Perspectives in clinical research, 3*(4), 149.

Bastian, H., Glasziou, P., & Chalmers, I. (2010). Seventy-five trials and eleven systematic reviews a day: how will we ever keep up? *PLoS medicine, 7*(9), e1000326.

Bates, D., Maechler, M., & Bolker, B. (2012). lme4: Linear mixed-effects models using S4 classes.

Bates, D. M. (2010). lme4: Mixed-effects modeling with R. *URL http://lme4. r-forge. r-project. org/book*.

Berwick, D. M. (1989a). Continuous improvement as an ideal in health care. *The New England Journal of Medicine, 320*(1), 53.

Berwick, D. M. (1989b). Health services research and quality of care: assignments for the 1990s. *Medical Care*, 763-771.

Berwick, D. M. (1991). Controlling variation in health care: a consultation from Walter Shewhart. *Medical Care, 29*(12), 1212-1225.

Berwick, D. M. (2002). A user's manual for the IOM's 'Quality Chasm'report. *Health Affairs, 21*(3), 80-90.

Berwick, D. M., Godfrey, A. B., Roessner, J., Plsek, P. E., & Garvin, D. A. (1990). *Curing health care: new strategies for quality improvement; a report on the national demonstration project on quality improvement in health care*: Jossey-Bass Publishers.

Best, M., & Neuhauser, D. (2004). Avedis Donabedian: father of quality assurance and poet. *Quality and Safety in Health Care, 13*(6), 472-473.

Best, M., & Neuhauser, D. (2004). Ignaz Semmelweis and the birth of infection control. *Quality and Safety in Health Care, 13*(3), 233-234.

Best, M., & Neuhauser, D. (2006). Walter A Shewhart, 1924, and the Hawthorne factory. *Quality and Safety in Health Care, 15*(2), 142-143.

Betancourt, J. R., & Maina, A. W. (2004). The Institute of Medicine report" Unequal Treatment": implications for academic health centers. *The Mount Sinai journal of medicine, New York, 71*(5), 314-321.

Bice, T. W. (1980). Social Science and Health Services Research: Contributions to Public Policy. *The Milbank Memorial Fund Quarterly. Health and Society, 58*(2), 173-200. doi: 10.2307/3349711

Bland, J. M. (2004). Cluster randomised trials in the medical literature: two bibliometric surveys. *BMC Medical Research Methodology, 4*(1), 21.

Bonnie, S., & Martin, R. (1998). Understanding controlled trials: Why are randomised controlled trials important? *BMJ, 316*.

Box, G., Bisgaard, S., & Fung, C. (1988). An explanation and critique of Taguchi's contributions to quality engineering. *Quality and reliability engineering international, 4*(2), 123-131.

Box, G. E. (1989). The RA Fisher memorial lecture, 1988: quality improvement: an expanding domain for the application of scientific method. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences, 327*(1596), 617-630.

Box, G. E., Hunter, J. S., & Hunter, W. G. (2005). *Statistics for experimenters*: Wiley New York.

Brennan, P. F., & Safran, C. (2004). Patient Safety: Remember who it's really for. *International Journal of Medical Informatics, 73*(7), 547-550.

Brennan, T. A. (2002). Physicians' professional responsibility to improve the quality of care. *Academic Medicine, 77*(10), 973-980.

Brennan, T. A., Leape, L. L., Laird, N. M., Hebert, L., Localio, A. R., Lawthers, A. G., . . . Hiatt, H. H. (1991). Incidence of adverse events and negligence in hospitalized patients: results of the Harvard Medical Practice Study I. *New England Journal of Medicine, 324*(6), 370-376.

Brierley, G., Brabyn, S., Torgerson, D., & Watson, J. (2012). Bias in recruitment to cluster randomized trials: a review of recent publications. *Journal of Evaluation in Clinical Practice, 18*(4), 878-886. doi: 10.1111/j.1365-2753.2011.01700.x

Campbell, M., Fitzpatrick, R., Haines, A., Kinmonth, A. L., & et al. (2000). Framework for design and evaluation of complex interventions to improve health. *British Medical Journal, 321*(7262), 694-696.

Campbell, M. J., Donner, A., & Klar, N. (2007). Developments in cluster randomized trials and Statistics in Medicine. *Statistics in Medicine, 26*(1), 2-19. doi: 10.1002/sim.2731

Campbell, M. K., Elbourne, D. R., & Altman, D. G. (2004). CONSORT statement: extension to cluster randomised trials. *BMJ, 328*(7441), 702-708. doi: 10.1136/bmj.328.7441.702

Campbell, N. C., Murray, E., Darbyshire, J., Emery, J., Farmer, A., Griffiths, F., . . . Kinmonth, A. L. (2007). Designing and evaluating complex interventions to improve health care. *BMJ, 334*(7591), 455-459. doi: 10.1136/bmj.39108.379965.BE

Carnes, M., Devine, P. G., Baier Manwell, L., Byars-Winston, A., Fine, E., Ford, C. E., . . . Sheridan, J. (2014). Effect of an Intervention to Break the Gender Bias Habit: A Cluster Randomized, Controlled Trial. *Academic Medicine, AcadMed-D-13-01441R2*.

Carnes, M., Devine, P. G., Isaac, C., Manwell, L. B., Ford, C. E., Byars-Winston, A., . . . Sheridan, J. (2012). Promoting institutional change through bias literacy. *Journal of Diversity in Higher Education, 5*(2), 63-77. doi: 10.1037/a002812810.1037/a0028128.supp (Supplemental)

Carter, B. L., Bergus, G. R., Dawson, J. D., Farris, K. B., Doucette, W. R., Chrischilles, E. A., & Hartz, A. J. (2008). A cluster randomized trial to evaluate physician/pharmacist collaboration to improve blood pressure control. *J Clin Hypertens (Greenwich), 10*(4), 260-271.

Chalmers, I., Dukan, E., Podolsky, S., & Smith, G. D. (2012). The advent of fair treatment allocation schedules in clinical trials during the 19th and early 20th centuries. *Journal of the Royal Society of Medicine, 105*(5), 221-227.

Chassin, M. R., & Galvin, R. W. (1998). The urgent need to improve health care quality: Institute of Medicine National Roundtable on Health Care Quality. *Jama, 280*(11), 1000-1005.

Chassin, M. R., & Loeb, J. M. (2011). The ongoing quality improvement journey: next stop, high reliability. *Health Affairs, 30*(4), 559-568.

Cornfield, J. (1978). Randomization by group: a formal analysis. *Am J Epidemiol, 108*(2), 100-102.

Craig, P., Dieppe, P., Macintyre, S., Michie, S., Nazareth, I., & Petticrew, M. (2008a). Developing and evaluating complex interventions: the new Medical Research Council guidance. *BMJ: British Medical Journal, 337*.

Craig, P., Dieppe, P., Macintyre, S., Michie, S., Nazareth, I., & Petticrew, M. (2008b). Developing and evaluating complex interventions: the new Medical Research Council guidance. *BMJ, 337*.

Cuckler, G. A., Sisko, A. M., Keehan, S. P., Smith, S. D., Madison, A. J., Poisal, J. A., . . . Stone, D. A. (2013). National health expenditure projections, 2012–22: slow growth until coverage expands and economy improves. *Health Affairs, 32*(10), 1820-1831.

D'Agostino, R. B. (1998). Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine, 17*(19), 2265-2281. doi: 10.1002/(sici)1097-0258(19981015)17:19<2265::aid-sim918>3.0.co;2-b

d'Agostino, R. B. (1998). Tutorial in biostatistics: propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Stat Med, 17*(19), 2265-2281.

Davim, J. P. (2012). *Statistical and Computational Techniques in Manufacturing*: Springer.

De Hoop, E., Teerenstra, S., Van Gaal, B. G. I., Moerbeek, M., & Borm, G. F. (2012). The "best balance" allocation led to optimal balance in cluster-controlled trials. *Journal of Clinical Epidemiology, 65*(2), 132-137. doi: http://dx.doi.org/10.1016/j.jclinepi.2011.05.006

De Souza, L. B. (2009). Trends and approaches in lean healthcare. *Leadership in Health Services, 22*(2), 121-139.

Dehejia, R. H., & Wahba, S. (2002). Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and statistics, 84*(1), 151-161.

Deming, W. E. (1986). Out of the crisis. Cambridge, MA: Massachusetts Institute of Technology. *Center for Advanced Engineering Study, 6*.

Diehr, P. (1995). Breaking the matches in a paired t-test for community interventions when the number of pairs is small. *Stat Med, 14*(13), 1491-1504.

Donabedian, A. (1966). Evaluating the quality of medical care. *The Milbank memorial fund quarterly*, 166-206.

Donabedian, A. (1980). Methods for deriving criteria for assessing the quality of medical care. *Medical care review, 37*(7), 653.

Donabedian, A. (1988a). The end results of health care: Ernest Codman's contribution to quality assessment and beyond. *The Milbank Quarterly, 67*(2), 233-256; discussion 257-267.

Donabedian, A. (1988b). The quality of care: How can it be assessed? *Jama, 260*(12), 1743-1748.

Donner, A., Brown, K. S., & Brasher, P. (1990). A methodological review of non-therapeutic intervention trials employing cluster randomization, 1979–1989. *International Journal of Epidemiology, 19*(4), 795-800.

Donner, A., & Klar, N. (2002). *Design and Analysis of Cluster Randomization Trials in Health Research*. London: Arnold.

Donner, A., Taljaard, M., & Klar, N. (2007). The merits of breaking the matches: a cautionary tale. *Statistics in Medicine, 26*(9), 2036-2051.

Eagly, A. H., & Karau, S. J. (2002). Role congruity theory of prejudice toward female leaders. *Psychological Review, 109*(3), 573-598. doi: 10.1037/0033-295x.109.3.573

Eccles, M., Grimshaw, J., Campbell, M., & Ramsay, C. (2003). Research designs for studies evaluating the effectiveness of change and improvement strategies. *Quality and Safety in Health Care, 12*(1), 47-52.

Eldridge, S., Ashby, D., Bennett, C., Wakelin, M., & Feder, G. (2008). Internal and external validity of cluster randomised trials: systematic review of recent trials. *BMJ, 336*(7649), 876-880. doi: 10.1136/bmj.39517.495764.25

Eldridge, S., & Kerry, S. (2012). *A practical guide to cluster randomised trials in health services research* (Vol. 120): John Wiley & Sons.

Eldridge, S. M., Ashby, D., Feder, G. S., Rudnicka, A. R., & Ukoumunne, O. C. (2004). Lessons for cluster randomized trials in the twenty-first century: a systematic review of trials in primary care. *Clinical trials, 1*(1), 80-90.

Epstein, J. N., Langberg, J. M., Lichtenstein, P. K., Kolb, R., Altaye, M., & Simon, J. O. (2011). Use of an Internet Portal to Improve Community-Based Pediatric ADHD Care: A Cluster Randomized Trial. *Pediatrics, 128*(5), e1201-e1208. doi: 10.1542/peds.2011-0872

Fayers, P., Jordhøy, M., & Kaasa, S. (2002). Cluster-randomized trials. *Palliative medicine, 16*(1), 69-70.

Feldman, H. A., & McKinlay, S. M. (1994). Cohort versus cross-sectional design in large field trials: precision, sample size, and a unifying model. *Statistics in Medicine, 13*(1), 61-78.

Ferlie, E. B., & Shortell, S. M. (2001). Improving the quality of health care in the United Kingdom and the United States: a framework for change. *Milbank Q, 79*(2), 281-315.

Fisher, R. A. (1919). XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance. *Transactions of the Royal Society of Edinburgh, 52*(02), 399-433.

Fisher, R. A., & Yates, F. (1949). Statistical tables for biological, agricultural and medical research. *Statistical tables for biological, agricultural and medical research.*(Ed. 3.).

Fisher, S. R. A., Genetiker, S., Fisher, R. A., Genetician, S., Britain, G., & Généticien, S. (1970). *Statistical methods for research workers* (Vol. 14): Oliver and Boyd Edinburgh.

Fitzmaurice, G., & Molenberghs, G. (2009). Advances in longitudinal data analysis: an historical perspective. *Longitudinal Data Analysis*, 3-30.

Flynn, B. B., Schroeder, R. G., & Sakakibara, S. (1994). A framework for quality management research and an associated measurement instrument. *Journal of Operations management, 11*(4), 339-366.

Garrison, M. M., & Mangione-Smith, R. (2013). Cluster Randomized Trials for Health Care Quality Improvement Research. *Academic pediatrics, 13*(6), S31-S37.

Gavgani, A., Hodjati, M., Mohite, H., & Davies, C. (2002). Effect of insecticide-impregnated dog collars on incidence of zoonotic visceral leishmaniasis in Iranian children: a matchedcluster randomised trial. *The lancet, 360*(9330), 374-379.

Giraudeau, B., & Ravaud, P. (2009). Preventing Bias in Cluster Randomised Trials (Vol. 6, pp. 1-6): Public Library of Science.

Glynn, R. J., Brookhart, M. A., Stedman, M., Avorn, J., & Solomon, D. H. (2007). Design of cluster-randomized trials of quality improvement interventions aimed at medical care providers. *Medical Care, 45*(10), S38-S43.

Godfrey, A. B. (1999). *Juran's quality handbook*: McGraw Hill.

Greenfield, M. L. (2004). Of plagues, blights, and bloodletting: historical highlights of the randomized controlled trial *Society of Clinical Research Associates*.

Grimes, D. A. (1995). Clinical Research in Ancient Babylon: Methodologic Insights From the Book of Daniel. *Obstetrics & Gynecology, 86*(6), 1031-1034.

Grimshaw, J., Eccles, M., Marion, C., & Elbourne, D. (2005). Cluster Randomized Trials of Professional and Organizational Behavior Change Interventions in Health Care Settings. *Annals of the American Academy of Political and Social Science, 599*(ArticleType: research-article / Issue Title:

Place Randomized Trials: Experimental Tests of Public Policy / Full publication date: May, 2005 / Copyright © 2005 American Academy of Political and Social Science), 71-93. doi: 10.2307/25046095

Group, T. C. R. (1995a). Community Intervention Trial for Smoking Cessation (COMMIT): I. Cohort Results from a Four-Year Community Intervention. [Article]. *American Journal of Public Health, 85*(2), 183-192.

Group, T. C. R. (1995b). Community Intervention Trial for Smoking Cessation (COMMIT): II. Changes in Adult Cigarette Smoking Prevalence. [Article]. *American Journal of Public Health, 85*(2), 193-200.

Gustafson, D. H., Quanbeck, A. R., Robinson, J. M., Ford, J. H., Pulvermacher, A., French, M. T., . . . McCarty, D. (2013). Which elements of improvement collaboratives are most effective? A cluster-randomized trial. *Addiction, 108*(6), 1145-1157. doi: 10.1111/add.12117

Hall, N. S. (2007). R. A. Fisher and His Advocacy of Randomization. *Journal of the History of Biology, 40*(2), 295-325. doi: 10.2307/29737483

Hedeker, D., & Gibbons, R. D. (1997). Application of random-effects pattern-mixture models for missing data in longitudinal studies. *Psychological Methods, 2*(1), 64-78. doi: 10.1037/1082-989x.2.1.64

Heitjan, D. F., & Basu, S. (1996). Distinguishing "missing at random" and "missing completely at random". *The American Statistician, 50*(3), 207-213.

Higgins, J. P., & Green, S. (2008). *Cochrane handbook for systematic reviews of interventions* (Vol. 5): Wiley Online Library.

Hill, C., Corbett, C., & St Rose, A. (2010). *Why So Few? Women in Science, Technology, Engineering, and Mathematics*: ERIC.

Hsieh, F. Y., Lavori, P. W., Cohen, H. J., & Feussner, J. R. (2003). An Overview of Variance Inflation Factors for Sample-Size Calculation. *Evaluation & the Health Professions, 26*(3), 239-257. doi: 10.1177/0163278703255230

Isaac, C., Kaatz, A., Lee, B., & Carnes, M. (2012). An educational intervention designed to increase women's leadership self-efficacy. *CBE-Life Sciences Education, 11*(3), 307-322.

Ivers, N., Taljaard, M., Dixon, S., Bennett, C., McRae, A., Taleban, J., . . . Eccles, M. (2011). Impact of CONSORT extension for cluster randomised trials on quality of reporting and study methodology: review of random sample of 300 trials, 2000-8. *BMJ: British Medical Journal, 343*.

Ivers, N. M., Halperin, I. J., Barnsley, J., Grimshaw, J. M., Shah, B. R., Tu, K., . . . Zwarenstein, M. (2012). Allocation techniques for balance at baseline in cluster randomized trials: a methodological review. *Trials, 13*, 120. doi: 10.1186/1745-6215-13-120

Janis, I. L., Mann L. (1977). *Decision Making: A Psychological Analysis of Conflict, Choice, and Commitment*. London: Cassel and Collier Macmillan.

John, P. W. M., & John, P. W. (1971). *Statistical design and analysis of experiments*: SIAM.

Jones, B., & Nachtsheim, C. J. (2009). Split-plot designs: What, why, and how. *Journal of Quality Technology, 41*(4), 340-361.

Kaptchuk, T. J. (2001). The double-blind, randomized, placebo-controlled trial: Gold standard or golden calf? *Journal of Clinical Epidemiology, 54*(6), 541-549.

Kaur, R. (2013). Randomised controlled trials. *Indian Journal of Medical Specialities, 4*(2).

Klar, N., & Donner, A. (2001). Current and future challenges in the design and analysis of cluster randomization trials. *Statistics in Medicine, 20*(24), 3729-3740.

Klar, N., & Donner, A. (2004). The impact of EF Lindquist's 1940 text "Statistical Analysis in Educational Research" on cluster randomization. *James Lind Library Bulletin: Commentaries on the history of treatment evaluation*.

Koepsell, T. D. (1998). Epidemiologic issues in the design of community intervention trials. *Applied Epidemiology (eds Brownson RC & Petitti DB)*, 177-211.

Kohn, L. T., Corrigan, J. M., & Donaldson, M. S. (2000). *To err is human: building a safer health system* (Vol. 627): National Academies Press.

Kunz, R., Vist, G., & Oxman, A. (2007). Randomisation to protect against selection bias in healthcare trials. *Cochrane Database Syst Rev, 2*(2).

Laffel, G., & Blumenthal, D. (1989). The case for using industrial quality management science in health care organizations. *Jama, 262*(20), 2869-2873.

Laird, N. M. (1988). Missing data in longitudinal studies. *Statistics in Medicine, 7*(1-2), 305-315. doi: 10.1002/sim.4780070131

Laird, N. M., & Ware, J. H. (1982). Random-Effects Models for Longitudinal Data. *Biometrics, 38*(4), 963-974. doi: 10.2307/2529876

Landrigan, C. P., Parry, G. J., Bones, C. B., Hackbarth, A. D., Goldmann, D. A., & Sharek, P. J. (2010). Temporal trends in rates of patient harm resulting from medical care. *New England Journal of Medicine, 363*(22), 2124-2134.

Leape, L. L., & Berwick, D. M. (2005). Five years after To Err Is Human: what have we learned? *Jama, 293*(19), 2384-2390.

Leape, L. L., Brennan, T. A., Laird, N., Lawthers, A. G., Localio, A. R., Barnes, B. A., . . . Hiatt, H. (1991). The nature of adverse events in hospitalized patients: results of the Harvard Medical Practice Study II. *New England Journal of Medicine, 324*(6), 377-384.

Lewsey, J. (2004). Comparing completely and stratified randomized designs in cluster randomized trials when the stratifying factor is cluster size: a simulation study. *Statistics in Medicine, 23*(6), 897-905.

Leyrat, C., Caille, A., Donner, A., & Giraudeau, B. (2013). Propensity scores used for analysis of cluster randomized trials with selection bias: a simulation study. *Statistics in Medicine, 32*(19), 3357-3372. doi: 10.1002/sim.5795

Lind, J. (1757). *A Treatise on the Scurvy: In Three Parts, Containing an Inquiry Into the Nature, Causes, and Cure, of that Disease*: A. Millar.

Little, R. J. A. (1993). Pattern-Mixture Models for Multivariate Incomplete Data. *Journal of the American Statistical Association, 88*(421), 125-134. doi: 10.2307/2290705

Little, R. J. A. (1995). Modeling the Drop-Out Mechanism in Repeated-Measures Studies. *Journal of the American Statistical Association, 90*(431), 1112-1121. doi: 10.1080/01621459.1995.10476615

Loeb, M. B. (2002). Application of the development stages of a cluster randomized trial to a framework for evaluating complex health interventions. *BMC Health Services Research, 2*(1), 13.

Lohr, K. N. (2007). Emerging Methods in Comparative Effectiveness and Safety: Symposium Overview and Summary. *Medical Care, 45*(10), S5-S8. doi: 10.2307/40221549

Lohr, K. N., Brook, R. H., Kamberg, C. J., Goldberg, G. A., Leibowitz, A., Keesey, J., . . . Newhouse, J. P. (1986). Use of medical care in the RAND Health Insurance Experiment: diagnosis-and service-specific analyses in a randomized controlled trial. *Medical Care*, S1-S87.

Lohr, K. N., & Steinwachs, D. M. (2002). Health services research: an evolving definition of the field. *Health Services Research, 37*(1), 15.

Luce, J. M., Bindman, A. B., & Lee, P. R. (1994). A brief history of health care quality assessment and improvement in the United States. *Western journal of medicine, 160*(3), 263.

MacArthur, C., Winter, H., Bick, D., Knowles, H., Lilford, R., Henderson, C., . . . Gee, H. (2002). Effects of redesigned community postnatal care on womens' health 4 months after birth: a cluster randomised controlled trial. *The lancet, 359*(9304), 378-385.

Martin, D. C., Diehr, P., Perrin, E. B., & Koepsell, T. D. (1993). The effect of matching on the power of randomized community intervention studies. *Statistics in Medicine, 12*(3-4), 329-338.

Maxwell, R. J. (1984). Quality assessment in health. *British medical journal (Clinical research ed.), 288*(6428), 1470.

Mazor, K. M., Sabin, J. E., Boudreau, D., Goodman, M. J., Gurwitz, J. H., Herrinton, L. J., . . . Platt, R. (2007). Cluster Randomized Trials: Opportunities and Barriers Identified by Leaders of Eight Health Plans. *Medical Care, 45*(10), S29-S37. doi: 10.2307/40221554

McCarthy, T., & White, K. L. (2000). Origins of health services research. *Health Services Research, 35*(2), 375.

McDonald, L. (2001). Florence Nightingale and the early origins of evidence-based nursing. *Evidence based nursing, 4*(3), 68-69.

McEntegart, D. J. (2003). The pursuit of balance using stratified and dynamic randomization techniques: an overview. *Drug Information Journal, 37*(3), 293-308.

McGlynn, E. A., Asch, S. M., Adams, J., Keesey, J., Hicks, J., DeCristofaro, A., & Kerr, E. A. (2003). The quality of health care delivered to adults in the United States. *New England Journal of Medicine, 348*(26), 2635-2645.

McIntyre, D., Rogers, L., & Heier, E. J. (2001). Overview, history, and objectives of performance measurement. *Health Care Financing Review, 22*(3), 7-22.

Molenberghs, G., & Verbeke, G. (2005). *Models for discrete longitudinal data*: Springer.

Mullan, F. (2001). A founder of quality assessment encounters a troubled system firsthand. *Health Affairs, 20*(1), 137-141.

Murrary, D. M. (2002). *Design and Analysis of Group Randomised Trials*. New York: Oxford University Press.

Murray, D. M. (1998). *Design and analysis of group-randomized trials* (Vol. 29): Oxford University Press.

Murray, D. M., Varnell, S. P., & Blitstein, J. L. (2004). Design and analysis of group-randomized trials: a review of recent methodological developments. *American Journal of Public Health, 94*(3), 423.

Neuhauser, D. (2003). Florence Nightingale gets no respect: as a statistician that is. *Quality and Safety in Health Care, 12*(4), 317-317.

Nicolay, C., Purkayastha, S., Greenhalgh, A., Benn, J., Chaturvedi, S., Phillips, N., & Darzi, A. (2012). Systematic review of the application of quality improvement methodologies from the manufacturing industry to surgical healthcare. *British Journal of Surgery, 99*(3), 324-335.

O'Brien, K. L., Millar, E. V., Zell, E. R., Bronsdon, M., Weatherholtz, R., Reid, R., . . . Santosham, M. (2007). Effect of Pneumococcal Conjugate Vaccine on Nasopharyngeal Colonization among Immunized and Unimmunized Children in a Community-Randomized Trial. *The Journal of Infectious Diseases, 196*(8), 1211-1220. doi: 10.2307/30087154

Oakeshott, P., Kerry, S. M., & Williams, J. E. (1994). Randomized controlled trial of the effect of the Royal College of Radiologists' guidelines on general practitioners' referrals for radiographic examination. *British Journal of General Practice, 44*(382), 197-200.

Palmer, I. S. (1977). Florence Nightingale: reformer, reactionary, researcher. *Nursing research, 26*(2), 84-89.

Park, S., Palta, M., Shao, J., & Shen, L. (2002). Bias adjustment in analysing longitudinal data with informative missingness. *Statistics in Medicine, 21*(2), 277-291. doi: 10.1002/sim.992

Pauler, D. K., McCoy, S., & Moinpour, C. (2003). Pattern mixture models for longitudinal quality of life studies in advanced stage disease. *Statistics in Medicine, 22*(5), 795-809. doi: 10.1002/sim.1397

Perry, M., Faes, M., Reelick, M. F., Olde Rikkert, M. G. M., & Borm, G. F. (2010). Studywise minimization: A treatment allocation method that improves balance among treatment groups and makes allocation unpredictable. *Journal of Clinical Epidemiology, 63*(10), 1118-1122. doi: http://dx.doi.org/10.1016/j.jclinepi.2009.11.014

Pocock, S. J., & Simon, R. (1975). Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. *Biometrics*, 103-115.

Preece, D. A. (1990). R. A. Fisher and Experimental Design: A Review. *Biometrics, 46*(4), 925-935. doi: 10.2307/2532438

Prochaska, J. M., Prochaska, J. O., & Levesque, D. A. (2001). A transtheoretical approach to changing organizations. *Administration and Policy in Mental Health, 28*(4), 247-261.

Puffer, S., Torgerson, D. J., & Watson, J. (2003). Evidence for risk of bias in cluster randomised trials: Review of recent trials published in three general medical journals. *British Medical Journal, 327*(7418), 785-789.

Quanbeck, A. R., Gustafson, D. H., Ford, J. H., Pulvermacher, A., French, M. T., McConnell, K. J., & McCarty, D. (2011). Disseminating quality improvement: study protocol for a large cluster-randomized trial. *Implementation Science, 6*(1), 44.

Reid, P. P., Grossman, J. H., National Academies, P., Institute of, M., National Academy of, E., Fanjiang, G., & Compton, W. D. (2005). *Building a better delivery system : a new engineering/health care partnership*. Washington, D.C.: National Academies Press.

Roberts, J. S., Coale, J. G., & Redman, R. R. (1987). A history of the Joint Commission on Accreditation of Hospitals. *Jama, 258*(7), 936-940.

Rosenbaum, P. R., & Rubin, D. B. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika, 70*(1), 41-55. doi: 10.2307/2335942

Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association, 79*(387), 516-524.

Roy, A., Bhaumik, D. K., Aryal, S., & Gibbons, R. D. (2007). Sample Size Determination for Hierarchical Longitudinal Designs with Differential Attrition Rates. *Biometrics, 63*(3), 699-707. doi: 10.1111/j.1541-0420.2007.00769.x

Rubin, D. B. (1976). Inference and Missing Data. *Biometrika, 63*(3), 581-592. doi: 10.2307/2335739

Rubin, D. B. (1997). Estimating causal effects from large data sets using propensity scores. *Annals of Internal Medicine, 127*(8_Part_2), 757-763.

Rubin, D. B. (2001). Using propensity scores to help design observational studies: application to the tobacco litigation. *Health Services and Outcomes Research Methodology, 2*(3-4), 169-188.

Samsa, G., & Matchar, D. (2000). Can continuous quality improvement be assessed using randomized trials?[see comment]. *Health Services Research, 35*(3), 687.

Sarkar, D. (2008). *Lattice: multivariate data visualization with R*: Springer.

Schnipper, J. L., Hamann, C., Ndumele, C. D., & et al. (2009). Effect of an electronic medication reconciliation application and process redesign on potential adverse drug events: A cluster-randomized trial. *Archives of Internal Medicine, 169*(8), 771-780. doi: 10.1001/archinternmed.2009.51

Sequist, T. D., Fitzmaurice, G. M., Marshall, R., Shaykevich, S., Marston, A., Safran, D. G., & Ayanian, J. Z. (2010). Cultural Competency Training and Performance Reports to Improve Diabetes Care for Black PatientsA Cluster Randomized, Controlled Trial. *Annals of Internal Medicine, 152*(1), 40-46. doi: 10.7326/0003-4819-152-1-201001050-00009

Sequist, T. D., Gandhi, T. K., Karson, A. S., Fiskio, J. M., Bugbee, D., Sperling, M., . . . Bates, D. W. (2005). A Randomized Trial of Electronic Clinical Reminders to Improve Quality of Care for Diabetes and Coronary Artery Disease. *Journal of the American Medical Informatics Association, 12*(4), 431-437. doi: http://dx.doi.org/10.1197/jamia.M1788

Shackford, S. (2006). How then shall we change? *Journal of Trauma-Injury, Infection, and Critical Care, 60*(1), 1-7.

Shewhart, W. A. (1931). *Economic control of quality of manufactured product* (Vol. 509): ASQ Quality Press.

Simpson, J. M., Klar, N., & Donnor, A. (1995). Accounting for cluster randomization: a review of primary prevention trials, 1990 through 1993. *American Journal of Public Health, 85*(10), 1378-1383.

Slymen, D. J., & Hovell, M. F. (1997). Cluster versus individual randomization in adolescent tobacco and alcohol studies: illustrations for design decisions. *International Journal of Epidemiology, 26*(4), 765-771.

Spiegelhalter, D. J. (1999). Surgical Audit: Statistical Lessons from Nightingale and Codman. *Journal of the Royal Statistical Society. Series A (Statistics in Society), 162*(1), 45-58. doi: 10.2307/2680466

Stukel, T. A., Fisher, E. S., Wennberg, D. E., Alter, D. A., Gottlieb, D. J., & Vermeulen, M. J. (2007). Analysis of observational studies in the presence of treatment selection bias: effects of invasive cardiac management on AMI survival using propensity score and instrumental variable methods. *Jama, 297*(3), 278-285.

Team R Core. (2014). R: A language and environment for statistical computing.

Therneau, T. M. (1993). How many stratification factors are "too many" to use in a randomization plan? *Controlled Clinical Trials, 14*(2), 98-108.

Thompson, E. A. (1990). R. A. Fisher's Contributions to Genetical Statistics. *Biometrics, 46*(4), 905-914. doi: 10.2307/2532436

Torgerson, D. J. (2001). Contamination in trials: is cluster randomisation the answer? *BMJ: British Medical Journal, 322*(7282), 355.

Varkey, P., Karlapudi, S. P., & Bennet, K. E. (2008). Teaching quality improvement: a collaboration project between medicine and engineering. *American Journal of Medical Quality, 23*(4), 296-301.

Walsh, J. E. (1947). Concerning the Effect of Intraclass Correlation on Certain Significance Tests. *The Annals of Mathematical Statistics, 18*(1), 88-96. doi: 10.2307/2236105

West, K. P., Jr., Katz, J., Khatry, S. K., LeClerq, S. C., & et al. (1999). Double blind, cluster randomised trial of low dose supplementation with vitamin A or(beta) carotene on mortality related to pregnancy in Nepal. *British Medical Journal, 318*(7183), 570-575.

White, K. L. (2002). Jerry Morris and health services research in the USA. *International Journal of Epidemiology, 31*(3), 690-692.

Williamson, E., Morley, R., Lucas, A., & Carpenter, J. (2012). Propensity scores: From naïve enthusiasm to intuitive understanding. [Article]. *Statistical Methods in Medical Research, 21*(3), 273-293. doi: 10.1177/0962280210394483

Yates, F. (1964). Sir Ronald Fisher and the Design of Experiments. *Biometrics, 20*(2), 307-321. doi: 10.2307/2528399

Yealy, D. M., Auble, T. E., Stone, R. A., Lave, J. R., Meehan, T. P., Graff, L. G., . . . Fine, M. J. (2004). The emergency department community-acquired pneumonia trial: Methodology of a quality improvement intervention. *Annals of Emergency Medicine, 43*(6), 770-782. doi: http://dx.doi.org/10.1016/j.annemergmed.2003.09.013

Yong, J., & Wilkinson, A. (2001). Rethinking total quality management. [Article]. *Total Quality Management, 12*(2), 247-258. doi: 10.1080/09544120120011460

Yoshioka, A. (1998). Use of randomisation in the Medical Research Council's clinical trial of streptomycin in pulmonary tuberculosis in the 1940s. *BMJ: British Medical Journal, 317*(7167), 1220.

Zimmerman, D. R. (2003). Improving nursing home quality of care through outcomes data: the MDS quality indicators. *International journal of geriatric psychiatry, 18*(3), 250-257.

Zucker, D. M. (1990). An analysis of variance pitfall: The fixed effects analysis in a nested design. *Educational and Psychological Measurement*.

APPENDIX 1

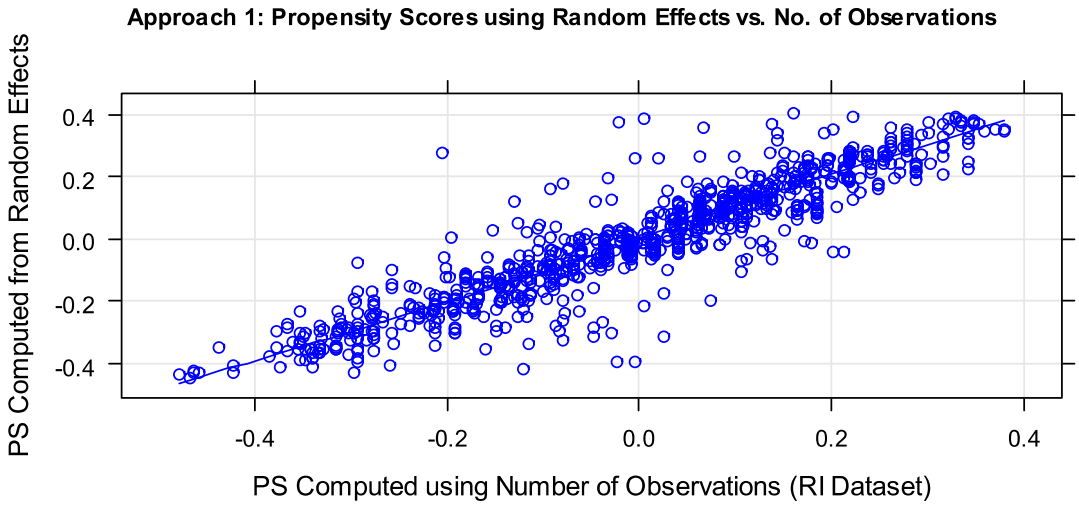

Figure A1.1: Linear relationship between propensity scores (PS) calculated using random effects compared to the those calculated using sum of observations in the RI attrition dataset

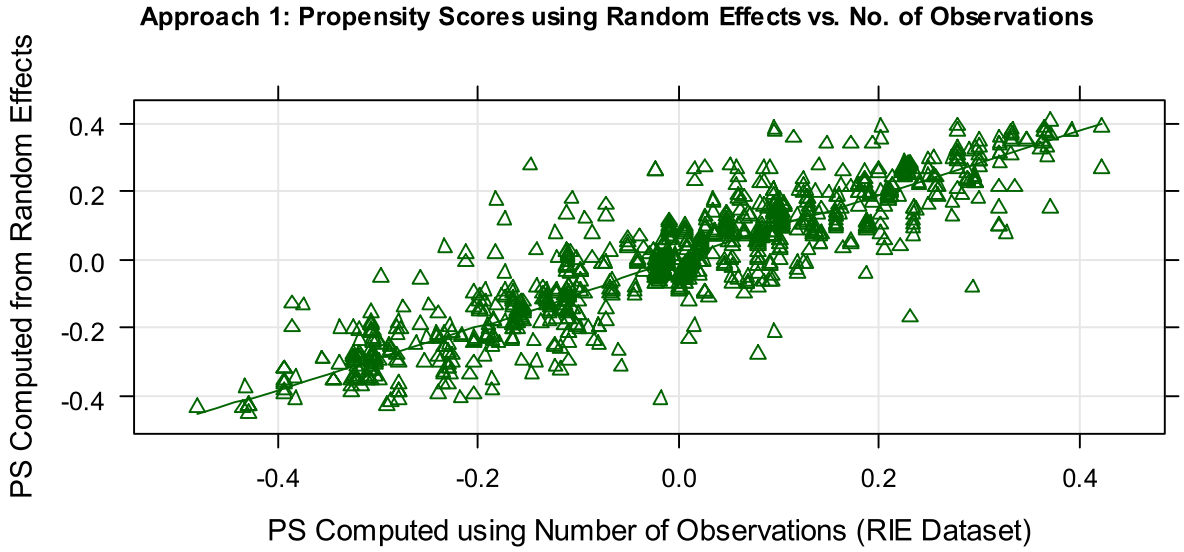

Figure A1.2: Linear relationship between propensity scores (PS) calculated using random effects compared to the those calculated using sum of observations in the RIE attrition dataset

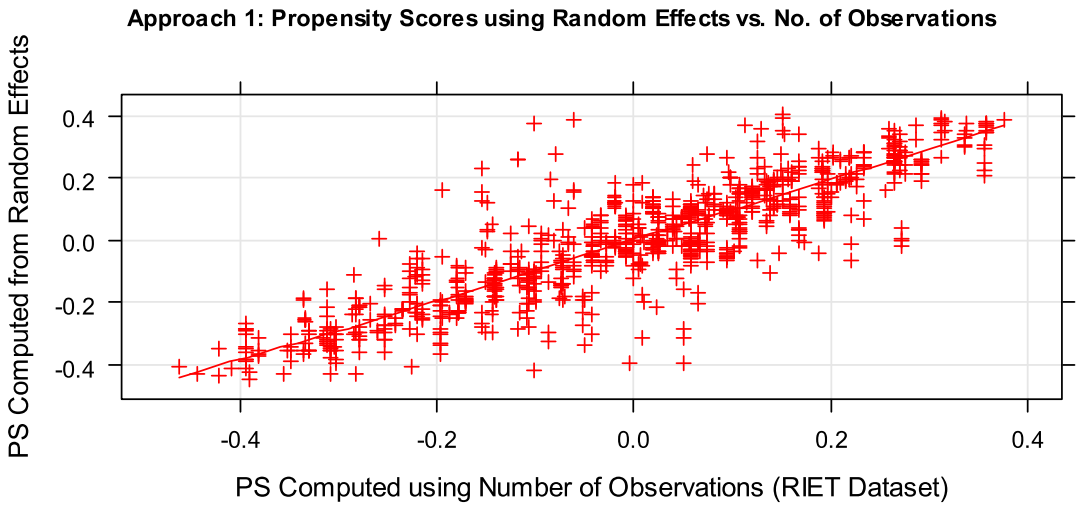

Figure A1.3: Linear relationship between propensity scores (PS) calculated using random effects compared to the those calculated using sum of observations in the RIET attrition dataset

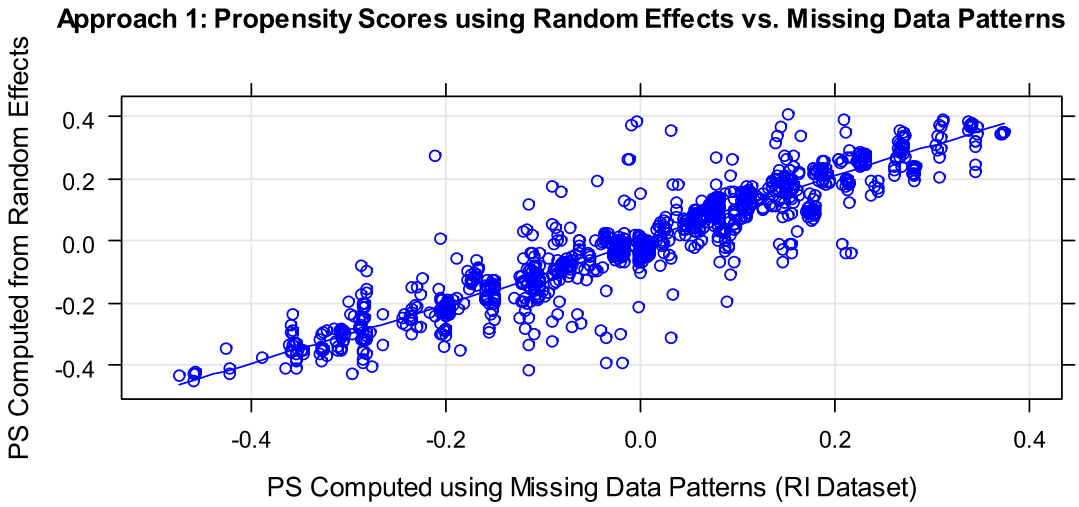

Figure A1.4: Linear relationship between propensity scores (PS) calculated using random effects compared to the those calculated using missing data patterns in the RI attrition dataset

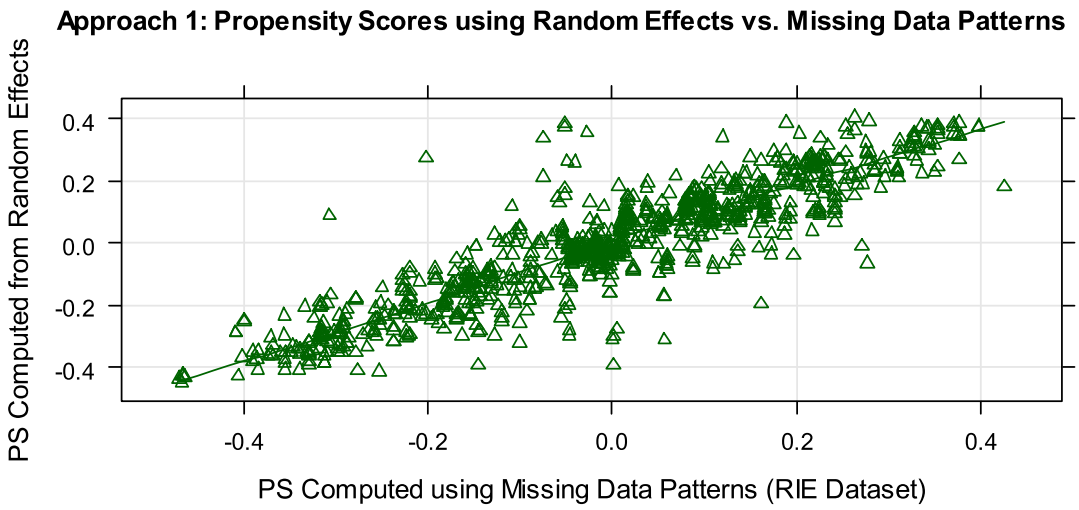

Figure A1.5: Linear relationship between propensity scores (PS) calculated using random effects compared to the those calculated using missing data patterns in the RIE attrition dataset

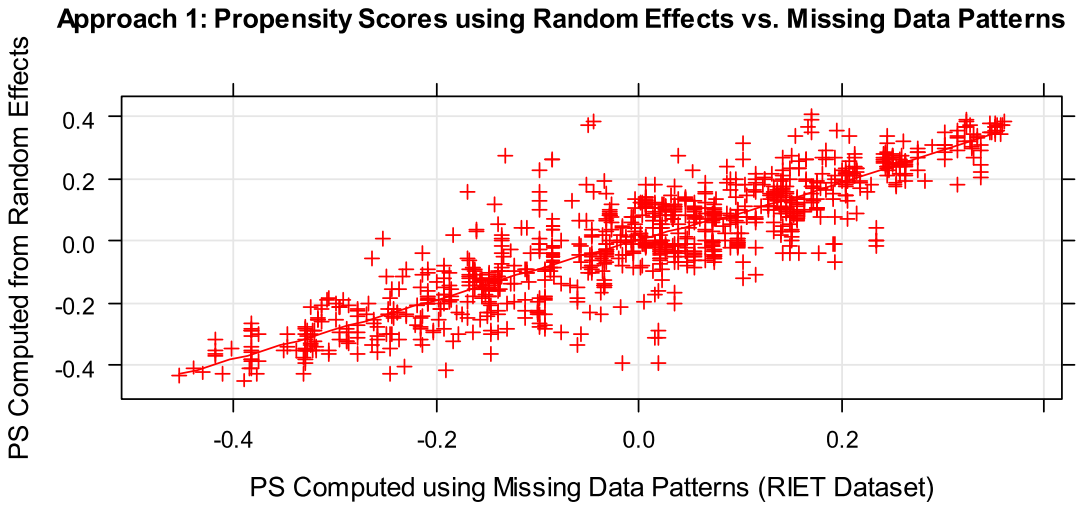

Figure A1.6: Linear relationship between propensity scores (PS) calculated using random effects compared to the those calculated using missing data patterns in the RIET attrition dataset