CRESST REPORT 832

TREATMENT CONFOUNDED MISSINGNESS: A COMPARISON OF METHODS FOR ADDRESSING CENSORED OR TRUNCATED DATA IN SCHOOL REFORM EVALUATIONS

SEPTEMBER, 2013

Jordan H. Rickles

Mark Hansen

Jia Wang

National Center for Research
on Evaluation, Standards, & Student Testing

UCLA | Graduate School of Education & Information Studies

Treatment Confounded Missingness: A Comparison of Methods for Addressing Censored or Truncated Data in School Reform Evaluations

CRESST Report 832

Jordan H. Rickles, Mark Hansen, and Jia Wang
CRESST/University of California, Los Angeles

September 2013

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
Center for the Study of Evaluation (CSE)
Graduate School of Education & Information Studies
University of California, Los Angeles
300 Charles E. Young Drive North
GSE&IS Bldg., Box 951522
Los Angeles, CA 90095-1522
(310) 206-1532

The work reported herein was supported under the grant number 52306 from the Bill and Melinda Gates Foundation with funding to the National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

The findings and opinions expressed here do not necessarily reflect the positions or policies of the Bill and Melinda Gates Foundation.

To cite from this report, please use the following as your APA reference: Rickles, J. H., Hansen, M., & Wang, W. (2013). *Treatment Confounded Missingness: A Comparison of Methods for Addressing Censored or Truncated Data in School Reform Evaluations* (CRESST Report 832). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

TABLE OF CONTENTS

TREATMENT CONFOUNDED MISSINGNESS: A COMPARISON OF METHODS FOR ADDRESSING CENSORED OR TRUNCATED DATA IN SCHOOL REFORM EVALUATIONS

Jordan H. Rickles, Mark Hansen, and Jia Wang
CRESST/ University of California, Los Angeles

Abstract

In this paper we examine ways to conceptualize and address potential bias that can arise when the mechanism for missing outcome data is at least partially associated with treatment assignment, an issue we refer to as treatment confounded missingness (TCM). In discussing TCM, we bring together concepts from the methodological literature on missing data, mediation, and principal stratification. We use a pair of simulation studies to demonstrate the main biasing properties of TCM and test different analytic approaches for estimating treatment effects given this missing data problem. We also demonstrate TCM and the different analytic approaches with empirical data from a study of a traditional high school that was converted to a charter school. The empirical illustration highlights the need to investigate possible TCM bias in high school intervention evaluations, where there is often an interest in studying the effects of an intervention or reform on both school persistence and academic achievement.

Introduction

Often, evaluations of educational interventions such as small learning communities, curriculum reform, and charter schools seek to estimate the causal effect of these interventions on student outcomes having to do with school engagement, progress, and academic achievement over multiple years. For example, a national evaluation of charter schools (Gleason, Clark, Tuttle, & Dwoyer, 2010) looked at student achievement over two years, and also examined outcomes pertaining to attendance, effort in school, behavior, and well-being. Even when such evaluations are based on the most rigorous experimental or quasi-experimental designs, the validity of effect estimates is threatened when missing data manifests for some outcomes. In this paper, we examine ways to conceptualize and address potential bias that can arise when the mechanism for missing outcome data is at least partially associated with treatment assignment, an issue we refer to as treatment confounded missingness (TCM). In discussing TCM, we bring together concepts from the methodological literature on missing data, mediation, and principal stratification.

Typically, researchers faced with missing outcome data restrict their analysis to units with observed data or employ standard data imputation methods that require explicit assumptions

about the missing data mechanism (Schafer & Graham, 2002). Two conditions can complicate these traditional missing data approaches, however. First, if observation of a primary outcome (e.g., academic achievement) is determined by an intermediary outcome (e.g., school dropout), restricting an analysis to units with observed outcome data implies conditioning on a post-treatment variable. This post-treatment conditioning can bias treatment effect estimates if unobserved factors influence both the intermediate outcome and the primary outcome (Pearl, 2009). Second, if the primary outcome is only defined for units that meet a certain threshold based on the intermediate outcome, the analysis should be restricted to the population for which treatment effects are defined. In medical research, for example, a treatment effect on patient quality of life is only defined for patients who are still living at the time of the post-treatment assessment, so the quality of life analysis targets the "survivor average causal effect" (SACE) rather than the population-wide average causal effect (Zhang & Rubin, 2003). When both of these conditions are present, the researcher faces a conundrum: estimate treatment effects with the observed data at the risk of inducing bias, only focus on outcomes fully defined for all individuals in the target population, or redefine the analysis plan to identify and exclude individuals for whom the treatment effect is undefined. Within the past decade, principal stratification (Frangakis & Rubin, 2002) has been applied to studies with TCM (McConnell, Stuart, & Devaney, 2008; Zhang & Rubin, 2003). The approach, as well as general acknowledgement of the TCM problem, is rarely addressed in educational evaluations, however.

To raise further awareness about possible TCM issues in educational evaluations, and school reform evaluations in particular, we present in this paper an empirical illustration and simulation study results with the following three objectives:

1. Demonstrate the conditions under which TCM will bias treatment effect estimates;
2. Demonstrate the principal stratification (PS) approach as it applies to a charter school evaluation; and
3. Compare the PS approach to alternative methods for estimating treatment effects given TCM.

The paper is organized as follows. In the next section, we discuss the concepts behind TCM and principal stratification. We then present simulation results to describe sensitivity of treatment effect estimation under different conditions and analytic approaches. To demonstrate a situation where TCM may be a concern and some analytic approaches researchers might consider for addressing those concerns, we then illustrate different analytic approaches with empirical data from a charter school conversion study. We conclude with a discussion of implications for educational researchers.

Mediation, Principal Stratification and Missing Data

The concepts addressed in this paper are rooted in the potential outcomes framework for causal inference popularized by Rubin (1974) and the application of PS (Frangakis & Rubin, 2002) to the "truncation by death" problem (Zhang & Rubin, 2003). They also relate to methodological issues and concepts having to do with missing data and mediation analysis. In this section, we briefly review the key concepts from each of these research areas to formalize our definition of TCM and potential issues that can arise in treatment effect estimation.

Potential Outcomes Framework

Under the potential outcomes framework, causal effects are defined at the individual unit level (e.g., a student) as the difference between the unit's outcome under a treatment condition and the same unit's outcome in the absence of the treatment (for simplicity, assume two treatment conditions):

$$\delta_i = Y(1)_i - Y(0)_i, \tag{1}$$

where unit i has a potential outcome $Y(1)_i$ if assigned to a treatment condition and a potential outcome $Y(0)_i$ if assigned to an alternative condition (i.e., the control condition). However, one can only observe a single potential outcome for the same unit under the same exact conditions (e.g., at the same point in time), a reality often referred to as the "fundamental problem of causal inference" (Holland, 1986). For example, for a student assigned to a treatment group (D_i=1), we will observe $Y(1)_i$, and $Y(0)_i$ only exists under an unobserved counterfactual condition. Conversely, for a student assigned to a control group (D_i=0), we will observe $Y(0)_i$, and $Y(1)_i$ only exists under an unobserved counterfactual condition.

While unobserved potential outcomes preclude us from estimating causal effects for an individual unit, we can estimate average treatment effects across units in a given sample based on the mean observed outcome for each treatment group:

$$\hat{\delta} = [\bar{Y}(1) | \, D = 1] - [\bar{Y}(0) | \, D = 0] \tag{2}$$

Under random assignment to treatment conditions, the above group-mean difference will be an unbiased estimate of the average treatment effect ($\hat{\delta}$). When the assignment to treatment conditions is not random, one can estimate an average treatment effect conditional on observed pretreatment covariates ($\mathbf{X}$):

$$\hat{\delta} = [\bar{Y}(1) | \, D = 1, \mathbf{X}] - [\bar{Y}(0) | \, D = 0, \mathbf{X}] \tag{3}$$

For the above conditional group-mean difference to be an unbiased estimate of the average treatment effect, the vector of conditioning covariates must include all the pretreatment factors that partially determine both treatment assignment and the potential outcomes. Since one cannot determine from the data alone whether all the important confounding factors are included in **X**, one must invoke an assumption of ignorable treatment assignment (Rosenbaum & Rubin, 1983), or selection on observables (Heckman & Hotz, 1989). Breakdowns in this assumption result in what is commonly referred to as selection bias (Shadish, Cook, & Campbell, 2002).

Mediation

Often, researchers are not just interested in a treatment's overall average effect on an outcome, but in whether the treatment effect is mediated by certain intermediary conditions. Much research has focused on defining and disentangling a treatment's direct and mediated, or indirect, effects (Bollen, 1987; Holland, 1988; Jo, 2008; Judd & Kenny, 1981). A textbook example of mediation is depicted in Figure 1a, where the effect of *D* on *Y* is mediated by *R*. Under a linearity assumption, the direct effect of *D* on *Y* is represented by *c*, and the indirect effect of *D* on *Y* via *R* is represented by $a \times b$.

Yet the estimation of direct and indirect effects is rarely as straightforward as the textbook example (Green, Ha, & Bullock, 2010). Even if treatment conditions, *D*, are randomly assigned, the identification of the direct and indirect effect can be confounded by factors that influence both *R* and *Y*, and failure to account for these factors will result in biased estimates of *b*. Such a case is depicted in Figure 1b, where an unobserved factor, *U*, is introduced. Pearl (2009) showed that conditioning on a factor like *R* that falls on the causal path from *D* to *Y* can result in "collider bias" if another factor jointly causes *R* and *Y*. In the case depicted in Figure 1b, conditioning on *R* induces an association between *D* and *U*, confounding an estimate of *D*'s effect on *Y* given *R*. Thus, getting unbiased estimates of mediated effects requires an assumption of sequential ignorability (Imai, Keele, & Yamamoto, 2010), where treatment assignment and the mediator are independent of the potential outcomes given measured pre-treatment covariates.

(a) Standard mediation analysis (b) Mediation analysis with confounding

Figure 1. Path diagrams illustrating mediation and confounding due to collider bias from an unobserved factor (U).

Principal Stratification

Within the potential outcomes framework, mediation analysis has been handled through principal stratification (PS). The PS approach has been used to address estimation problems arising from treatment noncompliance (Angrist, Imbens, & Rubin, 1996; Barnard, Frangakis, Hill, & Rubin, 2003) and, more recently, mediation (Jo, Stuart, MacKinnon, & Vinokur, 2011). The PS approach proposes that each unit belongs to one of four classes prior to treatment assignment (assuming, for simplicity, binary treatment assignment and binary mediator condition). These classes identify which intermediate event and potential outcome we observe, given the treatment assignment condition. For example, if students are assigned to a charter school (D=1) or traditional school (D=0), where treatment assignment can affect learning (Y) directly and indirectly based on whether the student remains in school (R=1) or leaves school (R=0), then students fall into one of four latent principal strata as defined in Table 1:

- Always Stayer (AS) – will remain in school regardless of assignment to a charter or traditional school;
- Encourager (EN) – will only remain in school if assigned to a charter school;
- Discourager (DS) – will only remain in school if assigned to a traditional school;
- Always Leaver (AL) – will leave school regardless of assignment to a charter or traditional school.

Table 1

Definition of Principal Strata, Associated Intermediate Event and Potential Outcomes

	Intermediate event (R)		Potential outcomes (Y)	
Principal strata	If D=1	If D=0	If D=1	If D=0
Always Stayer (AS)	R=1	R=1	$Y(D$=1,R=1)	$Y(D$=0,R=1)
Encourager (EN)	R=1	R=0	$Y(D$=1,R=1)	$Y(D$=0,R=0)
Discourager (DS)	R=0	R=1	$Y(D$=1,R=0)	$Y(D$=0,R=1)
Always Leaver (AL)	R=0	R=0	$Y(D$=1,R=0)	$Y(D$=0,R=0)

Notes: D=1 if assigned to treatment group and D=0 if assigned to control group.

Given the above principal strata and potential outcomes outlined in Table 1, the average direct causal effect of treatment can be defined as:

$$(\pi^{AS}/(\pi^{AS}+\pi^{AL}))\times(\mathrm{E}[Y(D=1,R=1)]-\mathrm{E}[Y(D=0,R=1)])+ \quad (4)$$

$$(\pi^{AL}/(\pi^{AS}+\pi^{AL}))\times(\mathrm{E}[Y(D=1,R=0)]-\mathrm{E}[Y(D=0,R=0)]),$$

where π^{AS} is the proportion of Always Stayers in the population and π^{AL} is the proportion of Always Leavers in the population. Similarly, the average indirect causal effect of treatment can be defined as:

$$(\pi^{EN}/(\pi^{EN}+\pi^{DS}))\times(\mathrm{E}[Y(D=1,R=1)]-\mathrm{E}[Y(D=0,R=0)])+ \quad (5)$$

$$(\pi^{DS}/(\pi^{EN}+\pi^{DS}))\times(\mathrm{E}[Y(D=1,R=0)]-\mathrm{E}[Y(D=0,R=1)]),$$

where π^{EN} is the proportion of Encouragers in the population and π^{DS} is the proportion of Discouragers in the population. Equations 4 and 5 can be expanded to estimate average treatment effects conditional on observed pretreatment factors (**X**) when treatment assignment is not random.

Since principal strata membership and the counterfactual potential outcome are unobserved, these average causal effects cannot be directly estimated from the data. If, for example, one was interested in estimating the average direct treatment effect for students who stay in school (*R*=1), the true average causal effect is defined by the potential outcomes for the Always Stayer stratum. However, the observed mean outcome for treatment units (*D*=1) who stay in school is a mixture of Always Stayers and Encouragers,

$$\bar{Y}(D=1,R=1)=(p_{D=1}^{AS}/(p_{D=1}^{AS}+p_{D=1}^{EN}))\times\bar{Y}_{D=1}^{AS} \quad (6)$$

$$+(p_{D=1}^{EN}/(p_{D=1}^{AS}+p_{D=1}^{EN}))\times\bar{Y}_{D=1}^{EN},$$

while the observed mean outcome for control units (*D*=0) who stay in school is a mixture of Always Stayers and Discouragers,

$$\bar{Y}(D=0,R=1)=(p_{D=0}^{AS}/(p_{D=0}^{AS}+p_{D=0}^{DS}))\times\bar{Y}_{D=0}^{AS} \quad (7)$$

$$+(p_{D=0}^{EN}/(p_{D=0}^{AS}+p_{D=0}^{DS}))\times\bar{Y}_{D=0}^{DS},$$

where $p_{D=1}^{PS}$ represents the proportion of treatment units in a given principal stratum and $p_{D=0}^{PS}$ represents the proportion of control units in a given principal stratum.

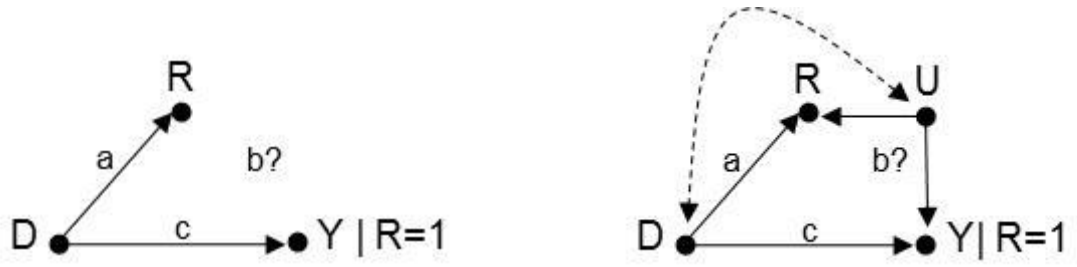
Principal stratification has primarily been used to address issues of noncompliance in randomized studies. In this application, two assumptions are typically invoked to facilitate estimation of the "complier average causal effect" (Angrist, Imbens, & Rubin, 1996): monotonicity and the exclusion restriction. Monotonicity assumes the effect of treatment assignment on the mediator is positive for all units. In other words, there is no Discourager stratum, or what is referred to as the Defier stratum in the noncompliance application. The exclusion restriction assumes treatment only affects the outcome through the mediator. In other

words, there is no direct effect of *D* on *Y*. While both of these assumptions are plausible in the noncompliance application, they are questionable in more general mediation applications. Recent work, however, has extended the principal stratification framework to more general mediation analysis applications (Gallop et al., 2009; Jo, 2008; Page, 2012).

Treatment Confounded Missingness

Researchers frequently encounter missing data. An overview of methods for handling missing data can be found in, for example, Allison (2001), Little and Rubin (2002) and Schafer and Graham (2002). In general, valid inferences with missing data hinge on the missingness, or response, mechanism that indicates whether a given variable is observed or not observed for a given unit (Rubin, 1976; Schafer & Graham, 2002). Based on the response mechanism, missing data can take one of three forms: missing completely at random (MCAR), missing at random (MAR), or not missing at random (NMAR). Data are MCAR when the response mechanism does not depend on the missing or observed data. Under such conditions, methods such as imputing missing data with unconditional means can result in unbiased estimates of population means. Data are MAR when the response mechanism depends on the observed data but not the missing data. Under MAR, methods such as imputing missing data with means conditional on the observed data can result in unbiased estimates of population means. Data are NMAR when the response mechanism depends on the missing data or unobserved factors. Under NMAR, imputing missing data will not result in unbiased estimates of population means.

We face an interesting challenge for treatment effect estimation when outcomes are not observed because of an intermediate, or mediating, event. For example, if we are interested in estimating the effect of a school reform on student test scores but scores are only observed for students who remain in school over a two-year period (R=1), then remaining in school acts as both a mediator and the missing data mechanism. Estimating the effect of school reform (D) on staying in school (R) is relatively straightforward, but ignoring the missing data when estimating the effect on achievement (Y) may result in biased inferences. First, because one must condition on R to observe Y, collider bias may confound estimates of D's effect on Y. To see this, consider the path diagrams in Figure 2. In Figure 2a, one can recover the direct effect of D on Y given R=1 because no other factors confound the relationship. Note that the indirect effect and the total effect cannot be estimated from the observed data because Y is never observed when R=0. In Figure 2b, however, conditioning on R induces an association between the unobserved factor (U) and D (represented by the dashed double-headed arrow), thus confounding the direct D to Y relationship. The distinction and introduction of bias is similar to missing data conditions under MAR versus NMAR assumptions, where missing Y values are MAR in Figure 2a but NMAR in Figure 2b.



(a) Analysis when observation of *Y* depends on *R*, but *Y* missing at random

(b) Analysis when observation of *Y* depends on *R*, but *Y* not missing at random

Figure 2. Path diagrams illustrating mediation and confounding when observation of the outcome depends on the mediating response.

From the perspective of principal stratification, this type of missing data process can bias treatment effect estimation because potential outcomes for certain strata are now missing. To address this missing data problem, researchers must determine whether the missing outcome data are best viewed as censored or truncated (McConnell et al., 2008). The distinction lies in whether the missing values could be observed but were censored due to attrition, or whether the outcome is undefined for specific units because the intermediate event rendered the outcome meaningless. In the first case, the estimand of interest is likely to be the average causal effect (ACE), and the censoring problem may be addressed with traditional missing data methods (Little & Rubin, 2002). The second case, where the outcome is not defined for some units, poses more complications and has been labeled a "truncation by death" problem (Zhang & Rubin, 2003). Here, the treatment effect of interest is restricted to those units in the principal stratum for which the outcome is defined under both treatment and control conditions (i.e., the Always Stayers). McConnell et al. (2008) refer to the estimand based on this stratum as the survivor average casual effect (SACE). To better see the distinction, in Table 2 we recast the potential outcomes for each principal strata, based on Table 1, given missing outcome data when R=0. From Table 2, one can see that only units belonging to the Always Stayer stratum have non-missing potential outcomes under treatment and under control conditions, whereas treatment effect estimation for the other strata are hindered by the missing data process. Furthermore, the observed outcome mean for units assigned to treatment will be a weighted average of the Always Stayers and Encouragers (see Equation 6) and the observed outcome mean for units assigned to control will be a weighted average of the Always Stayers and Discouragers. Therefore, if outcomes are truncated by *R*, the SACE should be the target estimand but may be biased by difficulties distinguishing between Always Stayers and Encouragers in the observed treatment group and Always Stayers and Discouragers in the observed control group. If outcomes are simply censored by *R*, however, one can target the ACE by imputing values for the missing outcomes inherent in

the other principal strata. As depicted in Figure 2, that approach will be sensitive to the MAR assumption.

Table 2

Definition of Principal Strata, Associated Intermediate Event and Potential Outcomes When Outcomes Are Missing for R=0.

Principal strata	Intermediate Event (R)		Potential Outcomes (Y)		Strata-specific Average Effect
	If D=1	If D=0	If D=1	If D=0	
Always Stayer (AS)	R=1	R=1	$Y(D$=1,R=1)	$Y(D$=0,R=1)	$Y(D$=1,R=1) - $Y(D$=0,R=1)
Encourager (EN)	R=1	R=0	$Y(D$=1,R=1)	*	*
Discourager (DS)	R=0	R=1	*	$Y(D$=0,R=1)	*
Always Leaver (AL)	R=0	R=0	*	*	*

Notes: *D=1* if assigned to treatment group and D=0 if assigned to control group.
* missing or undefined outcome/effect due to censoring or truncation.

We define the issue of truncated versus censored outcome data within a mediation analysis more generally as treatment confounded missingness (TCM). We have two main reasons for this more general label. First, it emphasizes the fact that the primary complication for treatment effect estimation given this type of missingness, censored or truncated, arises because treatment assignment (*D*) influences both the missing data mechanism (*R*) and the outcome of interest (*Y*). Second, it captures the notion that defining missing outcomes as censored or truncated can be a subjective decision. This may have implications for treatment effect estimation, particularly regarding whether the ACE or SACE is the estimand of interest. For example, in our motivating example, the evaluation of charter school effects was originally designed as a non-experimental study with propensity score matching to equate charter and traditional school students along pre-treatment covariates. After observing a significant charter school enrollment effect on staying in school, we had to determine how to estimate the effect of school type on other outcomes that may be sensitive to TCM (e.g., standardized test scores).

Prior work on the use of PS for estimating SACE focused on defining large sample bounds for treatment effects in randomized studies (Zhang & Rubin, 2003). In this paper, we are interested in the relevance of PS for non-experimental evaluations where one wants to obtain point estimates based on a finite sample. Key aspects of different analytic options for treatment effect point estimates are outlined in Table 3, with a more detailed description of each approach in Appendix A. In all these approaches, strong ignorability is assumed for the propensity score matching process, so potential outcomes for the matched treatment and control groups are independent of the pretreatment confounders if *Y* was fully observed. Given missing values in *Y*

when *R*=0, five general approaches to estimation of a treatment's direct effect on *Y* are outlined in Table 3. The first two approaches (1A and 1B) treat *Y* as censored and are designed to estimate the ACE. These two approaches either focus the analysis on cases with complete data (i.e., listwise deletion) or seek to impute the missing outcome values. The three other approaches treat *Y* as truncated and are designed to estimate the SACE, or the treatment effect for students in the Always Stayer stratum. The first approach that targets SACE (2A) re-runs the propensity score matching, restricting the matches to units with *Y* observed (and by definition *R*=1). This approach assumes all treatment units with *R*=1 are in the Always Stayer stratum (i.e., no Encouragers) and all control units with similar pre-treatment covariates as these treatment units are also in the Always Stayer stratum. The second approach (2B) uses the original matched sample, but restricts the analysis to matched treatment-control pairs that both have *Y* observed. This approach assumes the original matching process successfully paired units within the same principal strata, whereby the *R* outcome under the counterfactual condition can be inferred from the matched pair. The third approach (2C) also uses the original matched groups, but uses observed covariates to impute the *R* value under the counterfactual, rather than relying on matched pairs. Then, given the observed and imputed counterfactual *R* value, units in the Always Stayer strata are identified. Note that this approach is similar to the application of Bayesian techniques to classify units into principal strata (Page, 2012).

Table 3

Types of Methods For Estimating Treatment Effects Under Treatment Confounded Missingness For a Non-Experimental Study Using Propensity Score Matching

Nature of the unobserved outcome	Desired estimand	Analytical approach	Key assumptions
Y is censored	average causal effect (ACE)	1A. analysis of complete cases	Y is missing completely at random (MCAR)
		1B. multiple imputation of Y	Y is missing at random (MAR)
Y is truncated	survivor average causal effect (SACE)	2A. rematching of cases with observed Y	all subjects with Y observed belong to "always stayer" stratum (R_T=1,R_C=1)
		2B. analysis of "intact pairs"	matched subjects belong to same stratum
		2C. multiple imputation of retention (R) under alternative treatment assignment	R is missing at random (MAR)

In the following section we test the extent to which TCM is a concern under different data generating processes and whether certain analytic approaches are better suited for estimating treatment effects under TCM.

Simulation Study

To test the degree to which TCM can bias treatment effect estimation, we ran two Monte Carlo simulation studies. The first study examined the degree to which bias arises given a homogeneous null average treatment effect (i.e., treatment does not affect the outcome) and variation in both the proportion of units in the Encourager vs. Always Stayer strata and the relationship between an unobserved factor (*U*) and the mediating variable (*R*). The objective of this study was to better understand how strong the relationships between treatment (*D*) and *R* (which manifests through the size of the Encourager and Discourager strata) and between *U* and *R* have to be for TCM to become a concern. In addition, we sought to examine whether certain analytic approaches are more robust to changes in these relationships. In this study, the assumption of sequential ignorability breaks down when a relationship between *U* and *R* is present. The second study examined the degree to which bias arises given a heterogeneous average treatment effect and different conditions for the proportion of units in the Encourager vs. Always Stayer strata. The objective of this study is to better understand how TCM can be a

concern under heterogeneous treatment effects, and whether certain analytic approaches are better suited for isolating the SACE given heterogeneous treatment effects. In this study, the assumption of sequential ignorability holds, but the constant effect condition does not.

Both simulation studies are based on 1,000 Monte Carlo replications that produce a sample data set with a fixed sample size of 2,000 units. We used data from the empirical illustration (discussed in the following section) to guide the simulation data generating parameters. To simulate a non-experimental setting where the assumption of strong ignorability holds, treatment assignment for a given unit was based on a draw from a binomial distribution, where the probability of being assigned to the treatment condition (p_i) was a function of two "observed" covariates (*X1* & *X2*):

$$\ln\left(\frac{p_i}{1-p_i}\right) = -1.00 + 0.50\left(X1_i\right) + 0.25\left(X2_i\right).$$

Potential outcomes were a function of those two observed covariates as well as an "unobserved" covariate (*U*). In the first simulation study, the potential outcomes are generated based on a homogeneous null treatment effect:

$$Y(1)_i = Y(0)_i = 0.50\left(X1_i\right) + 0.25\left(X2_i\right) + 0.25\left(U_i\right) + e_i.$$

In the second simulation study, the relationship between *X1* and the potential outcome under treatment is stronger, thus creating a heterogeneous treatment effect that depends on *X1*:

$$\begin{aligned} Y(0)_i &= 0.50\left(X1_i\right) + 0.25\left(X2_i\right) + 0.25\left(U_i\right) + e_i \\ Y(1)_i &= Y(0)_i + \delta\left(X1_i\right), \end{aligned}$$

where δ takes on one of three values depending on the simulation condition:

- Small effect heterogeneity (δ = 0.25);
- Medium effect heterogeneity (δ = 0.50);
- Large effect heterogeneity (δ = 1.00).

In both simulation studies, e_i is drawn from a normal distribution with mean zero and standard deviation of 0.50.

Values for the three covariates are drawn from independent normal distributions with standard deviation of 1.00 and the mean for each distribution dependent on a unit's principal

stratum. In both simulation studies, the covariate population means for each stratum were based on the following matrix:

$$\begin{bmatrix} 0.50 & 0.50 & \upsilon 1 \\ 0.25 & 0.25 & \upsilon 2 \\ -0.25 & -0.25 & \upsilon 3 \\ -0.50 & -0.50 & \upsilon 4 \end{bmatrix}$$

where the rows correspond to the Always Stayer, Encourager, Discourager, and Never Stayer strata, respectively, and the columns correspond to *X1*, *X2*, and *U*, respectively. For *U*, in the first simulation study, principal strata mean differences vary across three conditions:

- No $U \rightarrow R$ effect (υ=[0 0 0 0])
- Moderate $U \rightarrow R$ effect (υ=[-0.25 -0.50 0.50 0.25])
- Large $U \rightarrow R$ effect (υ=[-0.50 -1.00 1.00 0.50])

In the second simulation study, only the No $U \rightarrow R$ effect condition is examined. The population means for *X1* and *X2* produce an overall sample where about one third of the units are assigned to treatment, but treatment selection is more prevalent among units in the Always Stayer and Encourager strata. Units in these strata also have higher average potential outcomes. Introducing between-strata variation in mean *U* values produces a $U \rightarrow R$ effect because the value of *R* depends on a unit's stratum membership and treatment assignment. Furthermore, censoring/truncation of the observed *Y* value depends on *R* (see Table 2).

Since data generation was designed to allow differences across principal strata, the proportion of units within each stratum is important. In both simulation studies, the proportion of units in the Discourager and Never Stayer strata were fixed at 0.10 and 0.15, respectively. These proportions provide a baseline for the overall mean of R and the $D \rightarrow R$ effect. We used three simulation conditions to look at performance under different proportions of units in the Always Stayer and Encourager strata:

- Small proportion of Encouragers ($\pi^{AS} = 0.65$, $\pi^{EN} = 0.10$)
- Medium proportion of Encouragers ($\pi^{AS} = 0.50$, $\pi^{EN} = 0.25$)
- Large proportion of Encouragers ($\pi^{AS} = 0.25$, $\pi^{EN} = 0.50$)

As the proportion of Encouragers increases, the $D \rightarrow R$ effect increases. Given between-strata heterogeneity in covariates and potential outcomes, if equations 4-7 hold, we hypothesize that bias from TCM will increase as the proportion of Encouragers relative to Always Stayers increases.

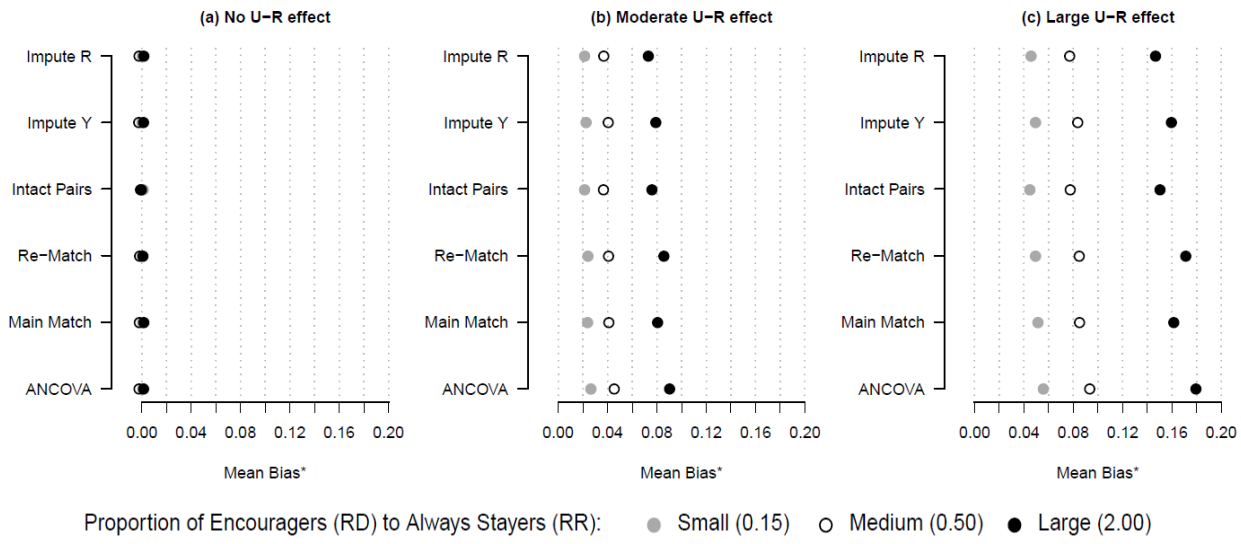
Under each condition and Monte Carlo replication we estimate average treatment effects based on six different analytic approaches: the five approaches discussed above (see Table 3) that start with a propensity score matched sample based on *X1* and *X2*, and one ANCOVA approach that conditions on *X1* and *X2* based on the full sample (details of each approach are in Appendix A). For simplicity, the matching-based approaches use 1-to-1 nearest neighbor matching and therefore target either the average treatment effect on the treated (ATT) or the survivor average treatment effect on the treated (SATT), rather than the overall average treatment effect or SACE. This distinction can be important in the second simulation study where treatment effect heterogeneity exists. For all analytic approaches, we assess performance based on three measures:

- Mean bias from the true population SATT;
- Root mean squared error (RMSE); and
- Type I error rate (study 1) or coverage rate (study 2).

In the first simulation study, the true population SATT is zero and we examine the Type I error rate to see whether TCM inflates the rate at which the null hypothesis is rejected when it should not be. In the second simulation study, the true population SATT depends on the *X1* mean for treatment units in the Always Stayer stratum (0.80 in the population under all conditions) and the size of δ, which differs across conditions. We examine the coverage rate to see whether the true SATT falls within ±2 standard errors of the estimated average effect at a more or less frequent rate under TCM. Results for both simulation studies are presented in Appendix B. Highlights of the results are discussed below.

Results from Simulation Study 1: Testing Collider Bias

Results from this simulation study show how treatment effect estimates can be biased under TCM because conditioning on a collider variable like *R* introduces confounding with *U*. As a result, bias increases as the effect of *D* on *R* increases (as determined by the relative size of the Encourager strata) and as the effect of *U* on *R* increases (see Figure 3). When *U* and *R* are independent, then all estimation methods can recover the true SATT, regardless of the size of the Encourager stratum (see Figure 3a). Similarly, for a given effect of U on R, the degree of bias depends on the size of the Encourager stratum. For example, for a moderate $U \rightarrow R$ effect, bias will only be about 0.02 with a relatively small Encourager population, but will be about four times larger with a relatively large Encourager population (see Figure 3b). None of the tested analytic approaches are able to account for this type of bias stemming from an unobserved covariate related to *R* and *Y*.



* Mean bias is displayed as positive deviations from the true average effect, even though mean bias was negative.

Figure 3. Mean treatment effect bias across simulation replications, by $U \rightarrow R$ effect, size of Encourager stratum, and analytic approach.

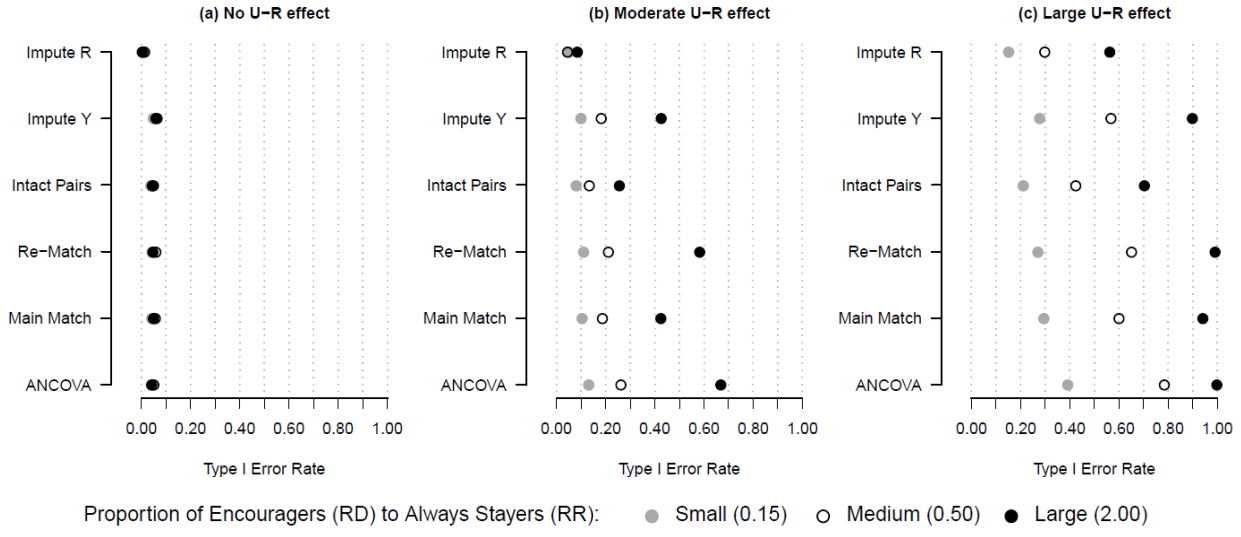


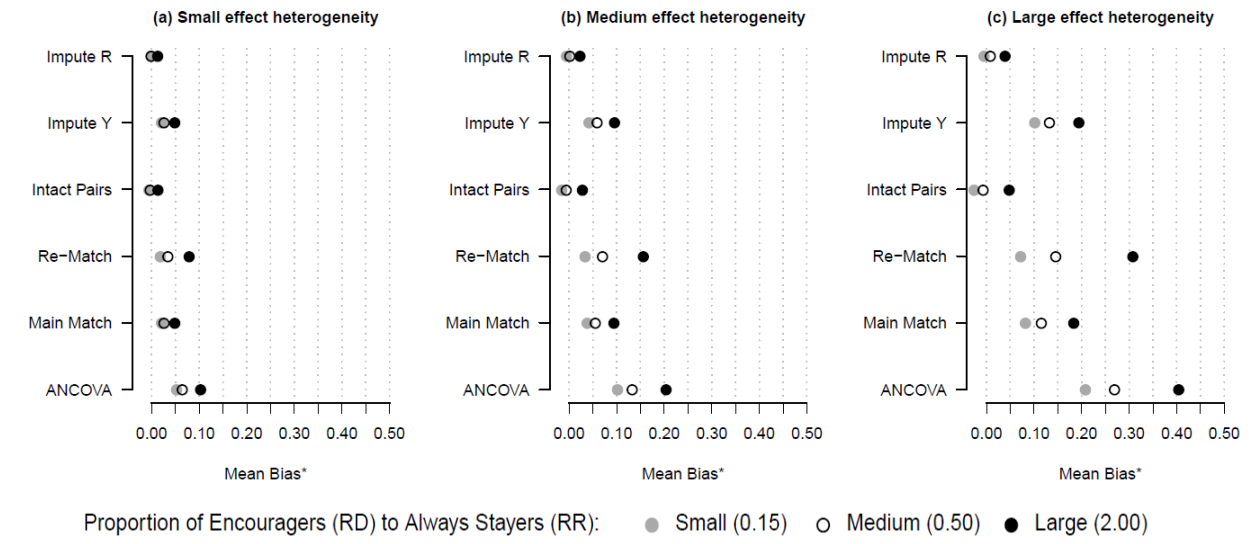
Figure 4. Type I error rates across simulation replications, by $U \rightarrow R$ effect, size of Encourager stratum, and analytic approach.

Type I error rates are sensitive to the analytic approach of choice, however, when bias is induced from a moderate to large $U \rightarrow R$ effect and a medium to large proportion of Encouragers (see Figure 4). Analytic approaches that explicitly target the SATT by treating missing Y values as truncated and focusing on the effect for treatment units in the Encourager stratum—by either imputing *R* or restricting the analysis to intact matched pairs—have lower Type I error rates than

the other methods. For example, the error rate under a moderate $U \rightarrow R$ effect and medium proportion of Encouragers (see Figure 4b) is less than 0.10 for the impute *R* method and just over 0.10 for the intact pairs method, but is almost 0.20 for the main match approach and well over 0.20 for the ANCOVA approach. This difference in performance is driven by the fact that the impute *R* and intact pairs methods restrict the analytic sample to units expected to be in the Encourager stratum. This smaller, and arguably more appropriate, sample size results in larger standard errors and therefore more frequent failure to reject the null hypothesis. It is important to note, however, that regardless of the analytic approach, the error rates are above the 0.05 level one would expect if TCM bias was not an issue.

Results from Simulation study 2: testing effect heterogeneity

Results from this simulation study show that even if one is not concerned with collider bias, treatment effect estimates can be biased under TCM if effect heterogeneity exists and the analytic approach does not target the appropriate estimand. As discussed above, in this simulation study, *U* and *R* are independent but the effect of *D* on *Y* depends on *X1*. Since Always Stayers have higher *X1* values, on average, than Encouragers, failure to isolate effect estimation to the Always Stayer stratum should downwardly bias estimates of the SATT. Mean bias is larger for analytic approaches that do not explicitly target the SATT and the Always Stayer stratum, particularly as the proportion of Encouragers increases (see Figure 5). When only a relatively small degree of effect heterogeneity exists (Figure 5a), the impute *R* and intact pairs approaches have little to no bias while bias with the other approaches is around 0.05 to 0.10. If a relatively large degree of effect heterogeneity exists (Figure 5b), bias is still minimal with the impute *R* and intact pairs approaches, while bias with the other approaches is anywhere between 0.10 and 0.40. Similar trends exist when looking at coverage rates (see Figure 6), with the impute *R* and intact pairs approaches providing coverage rates around 0.90 or higher, even under the large effect heterogeneity condition.



* Mean bias is displayed as positive deviations from the true average effect, even though mean bias was negative.

Figure 5. Mean treatment effect bias across simulation replications, by degree of effect heterogeneity, size of Encourager stratum, and analytic approach.

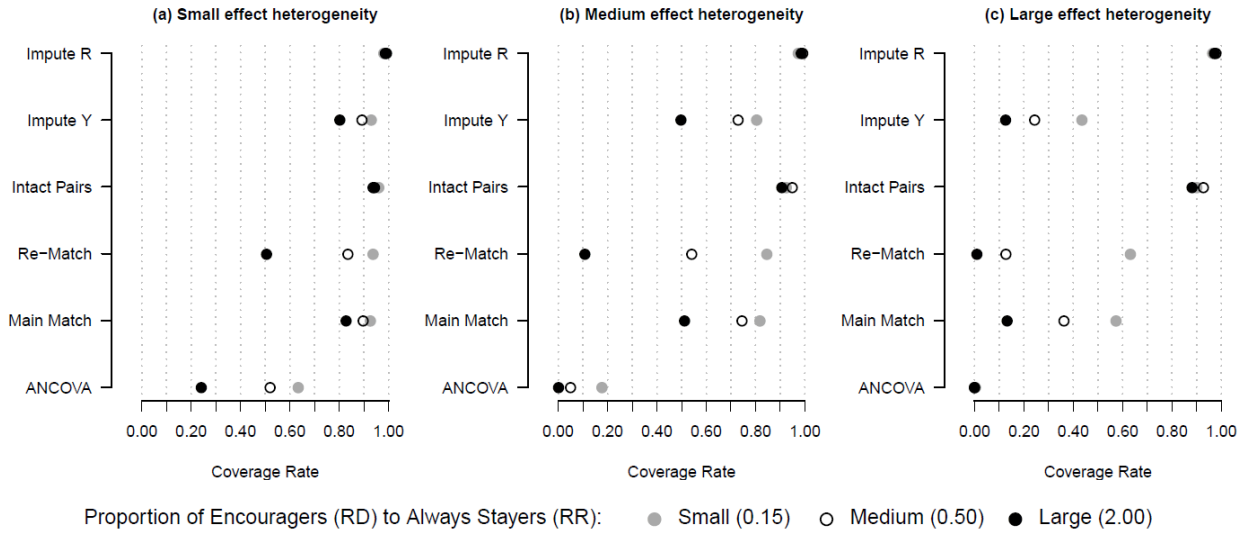


Figure 6. Coverage rates across simulation replications, by degree of effect heterogeneity, size of Encourager stratum, and analytic approach.

Empirical Illustration

The simulation study identified conditions under which TCM can result in biased treatment effect estimates and how some analytic approaches perform under those conditions. In this section we use an empirical example to illustrate TCM and the analytic approaches used for effect estimation. Our example was motivated by an evaluation of a charter school organization's

transformation of a traditional public high school (Herman, et al., 2012; Rickles, Wang & Herman, 2013). The primary analysis for the evaluation compared outcomes for two cohorts of students enrolled in the new charter school to a matched group of students who attended the same middle schools but neighboring traditional public high schools. A variety of high school outcomes were examined as students progressed from 9th grade through 12th grade, including: whether students remained in school from one year to the next; school attendance; how students performed on state standardized tests; and whether students took and passed the necessary courses for college eligibility. Overall, the evaluation found that students who attended the charter school, compared to observationally similar students attending demographically similar schools, were statistically more likely to stay at the same school for four years, take and pass key college preparatory courses, perform better on mathematics California Standards Tests (CST), pass the California High School Exit Exam (CAHSEE), and graduate within four years. The evaluation did not find a statistically significant effect on English language arts (ELA) CAHSEE scale scores; an issue we reexamine for this illustration. TCM may bias some of the findings because some outcomes like test performance and course completion are only observed for students who remain in school over a given period of time, and the evaluation found that the charter school students were more likely to stay in school.

For this illustration, we reanalyzed data on the student cohort entering 9th grade in 2008-09, and examined performance on the ELA portion of the CAHSEE taken in 10th grade.[1] Performance data were not available for students who did not take the test in their second year of high school (e.g., left the public school sample before test administration). The original matched sample included 1,162 students evenly distributed between treatment and control groups, while the sample of students with non-missing CAHSEE data included 502 treatment students and 446 control students. This indicates that 86% of the matched treatment students were tested but only 77% of the matched control students were tested. Descriptive pre-treatment characteristics of the treatment and control groups are presented in Table 4. Of those with observed outcome data, the raw mean difference between treatment and control groups is 0.03 standard deviations (95% C.I.: -0.10 to 0.16) and the addition of ANCOVA adjustment with the matched data produces a treatment effect estimate of 0.06 standard deviations (95% C.I.: -0.03 to 0.14). This represents an effect estimate that does not necessarily target the SATT and ignores TCM.

[1] Our analytic sample and methods for this illustrative example differed slightly from the methods used in the original analysis. For example, the original analysis used a combination of exact matching and nearest neighbor propensity score matching, but for the illustration we used optimal matching based solely on the propensity score. Therefore, the numbers and results do not necessarily align with what was published in the evaluation reports. Variables included in the optimal matching algorithm and in the covariate adjustment regression models are listed in Table 4, excluding the outcome missing data indicators and the outcome variables.

Zhang & Rubin (2003) provide a way to place bounds on the SATT given TCM. We applied this approach to our data and obtained a rather wide range of possible values for the true treatment effect (see Table 5). Without additional assumptions about the principal strata, the bounds range from -0.74 to 0.76. The bounds can be narrowed by invoking additional assumptions: monotonicity (i.e., no students in the Discourager stratum) and ranked averaged scores (i.e., the Always Stayers have higher average outcomes than the other principal strata). When both assumptions are invoked the bounds narrow to 0.03 to 0.23. For these data, however, the monotonicity assumption is not likely to hold. The assumption implies the charter school experience does not cause some students to leave school when they would have stayed in a traditional school. Since the charter school environment is not likely to benefit all students, this assumption is questionable. The rank averaged scores assumption, however, is more plausible. Unfortunately, just using this assumption still leaves us with rather wide bounds: -0.29 to 0.50.

Table 4

Description of Sample Used For the Empirical Illustration

	Full sample		Matched sample		Outcome observed	
Sample description	Treatment	Control	Treatment	Control	Treatment	Control
Number of students	581	1,467	581	581	502	446
% Female	48%	50%	48%	50%	49%	51%
Race/Ethnicity (%):						
Afr. Am./Black	25%	21%	25%	24%	25%	21%
Latino / Hispanic	74%	79%	74%	75%	75%	79%
Language classification (%):						
English only	34%	27%	34%	32%	33%	28%
English learner	36%	40%	36%	37%	36%	39%
Reclassified Eng. proficient	30%	34%	30%	31%	31%	33%
Parent education (%):						
HS graduate	24%	26%	24%	24%	25%	24%
Not HS graduate	28%	39%	28%	29%	28%	29%
Not available / missing	48%	34%	48%	48%	47%	47%
% free lunch	87%	84%	87%	88%	88%	88%
% student w/ disabilities	9%	7%	9%	9%	9%	8%
8th Grade attendance rate	0.94	0.94	0.94	0.94	0.94	0.95
8th Grade math test (%):						
Algebra 1	46%	73%	46%	48%	48%	50%
General Math	54%	27%	54%	52%	52%	50%
8th Grade test score:						
Math	0.05	-0.02	0.05	0.01	0.09	0.12

Sample description	Full sample		Matched sample		Outcome observed	
	Treatment	Control	Treatment	Control	Treatment	Control
ELA	0.13	-0.05	0.13	0.11	0.17	0.18
Missing outcome data (%):						
9th Grade ELA CST	5%	10%	5%	11%	2%	5%
10th Grade ELA CST	17%	19%	17%	22%	6%	4%
11th Grade ELA CST	30%	31%	30%	34%	21%	17%
ELA CAHSEE	14%	21%	14%	23%	0%	0%
Math CAHSEE	14%	20%	14%	23%	2%	1%
Outcome score (Z-score):						
9th Grade ELA	-0.01	0.01	-0.01	-0.06	0.03	0.02
10th Grade ELA	-0.04	0.02	-0.04	-0.06	-0.03	-0.04
11th Grade ELA	0.05	-0.02	0.05	-0.02	0.04	-0.01
Math CAHSEE	0.08	-0.03	0.08	-0.07	0.09	-0.06
ELA CAHSEE	-0.01	0.00	-0.01	-0.04	-0.01	-0.04

Table 5

Large sample bounds for survivor average treatment effect for the treated (SATT) based on Zhang & Rubin (2003) principal stratification method.

Assumptions	Lower bound	Upper bound
None	-0.74	0.79
Monotonicity	-0.19	0.23
Ranked average score	-0.29	0.50
Monotonicity & ranked average score	0.03	0.23

Since the bounds approach provides little direction for substantive conclusions—and may be inappropriate for finite samples—we can get alternative treatment effect point estimates based on the different analytic approaches discussed above. For this analysis, the average treatment effect estimates are fairly stable across the different analytic approaches (see Table 6). All the treatment effect estimates presented in Table 6 are based on units in the original matched sample, except the rematch approach (2A), and reflect regression-adjusted effect estimates to account for residual pretreatment covariate imbalance between the matched treatment and control groups. Treating the missing outcome values as censored and multiply imputing the missing values (approach 1B) resulted in only a slightly lower average effect estimate than the complete-case analysis based on the original matched sample (0.05 vs. 0.06). Both of these approaches target

the ATT rather than the SATT. However, results from the three approaches that target the SATT were also fairly similar, with the possible exception of the rematch approach (2A) that produced an average effect of -0.01. The intact pairs approach (2B) and isolating the Always Stayer stratum based on imputing the missing data mechanism (2C) resulted in average effect estimates of 0.03 and 0.06, respectively. Perhaps more importantly, there is a great deal of overlap in the confidence intervals for all these approaches, and with any of the approaches one would not reject the null hypothesis of no average treatment effect. Furthermore, it is worth noting that all the point estimates fall within the large sample bounds except for the rematch estimate when bounds are based on both the monotonicity and ranked average scores assumptions.

Table 6

Treatment Effect Estimates Based on Alternative Methods For Addressing Treatment Confounded Missingness

Nature of the Unobserved Outcome	Desired Estimand	Analytical Approach	N	Est	(SE)	95% C.I. (LB, UB)	
Y is censored	average causal effect (ATT)	1A. analysis of complete cases	948	0.06	(0.04)	-0.03	0.14
		1B. multiple imputation of Y	1,162	0.05	(0.04)	-0.03	0.14
Y is truncated	survivor average causal effect (SATT)	2A. rematching of cases with observed Y	1,004	-0.01	(0.05)	-0.12	0.10
		2B. analysis of "intact pairs"	770	0.03	(0.05)	-0.06	0.13
		2C. multiple imputation of retention (R) under alternative treatment assignment	785†	0.06	(0.05)	-0.04	0.16

Notes: all methods based on units in the original matched sample, except 2A, and include regression adjustment to account for residual pretreatment covariate imbalance between treatment and control groups.
†Average number of units in the "always taker" stratum across 1000 imputations.

One way to explore the degree to which effect estimates from the original matched sample might differ from the SATT is to look within principal strata as defined by the matched pairs. The simulation study results suggest that if effect heterogeneity exists, an analysis based on matched pairs will provide a less biased estimate of the SATT. Units were classified into one of the four principal strata based on whether the treatment and/or control units within each matched pair remained in school and took the CAHSEE. The treatment and control group outcome means for matched pairs within each of the assigned strata are presented in Table 7. Of the matched pairs, 66% had treatment and control units with non-missing outcome data. These pairs were

defined as the Always Stayers and comprise the target sample for the SATT. Restricting effect estimation to this stratum results in an unadjusted effect estimate of -0.03 (95% C.I.: -0.17 to 0.11), while including covariate adjustment with the intact pairs results in an average effect estimate of 0.03 (95% C.I.: -0.06 to 0.13). One can see how the raw mean difference for the Always Stayer stratum differs from the overall raw mean difference because of Encouragers (20% of the pairs) included in the overall treatment group mean and Discouragers (10% of the pairs) included in the overall control group mean. The mean outcome for the Encourager treatment units was higher than for Always Stayer treatment units, which produces an overall positive mean difference of 0.03 rather than the Always Stayer negative mean difference of -0.03. The Always Stayer matched pairs are not perfectly equated across all the pretreatment covariates, however. So adjusting for residual pretreatment differences among the Always Stayers pushes the estimated SATT to, coincidentally, 0.03 instead of -0.03.

Table 7

Observed Outcome Scores For Matched Pairs, By Principal Strata

Principal Strata	# of pairs	% of pairs	Treatment Group		Control Group		Mean Difference
			Mean	SD	Mean	SD	
(1) Always Takers: R=1, R=1	385	66%	-0.07	0.98	-0.04	1.00	-0.03
(2) Encouragers: R=1, R=0	117	20%	0.19	0.99	*	*	*
(3) Discouragers: R=0, R=1	61	10%	*	*	-0.03	1.12	*
(4) Never Takers: R=0, R=0	18	3%	*	*	*	*	*
Total	581	100%	-0.01	0.99	-0.04	1.02	0.03

* Missing/undefined

Discussion of Findings

Treatment confounded missingness can be seen as a special case of mediation analysis, where the mediator is also the missing data mechanism, or as a special case of missing data, where the missingness is partially determined by treatment assignment. The simulation study findings suggest that treatment effect bias from TCM is an increasing concern when three conditions arise. First, bias can increase as the proportion of treatment units in the Encourager stratum relative to the Always Stayer stratum increases. Second, bias can increase as the magnitude of an unobserved factor's effect on the mediating missing data mechanism increases (assuming the unobserved factor is also related to the primary outcome of interest). Third, bias can increase as between-strata heterogeneity in the treatment effect increases. While the proportion of Encouragers can be extrapolated from the size of the treatment effect on the

mediating missing data mechanism and assumptions about the size of the Discourager stratum, it is not possible to determine the extent of the other two biasing factors from the available data. Researchers are therefore left to make assumptions regarding sequential ignorability (i.e., no unobserved factors confounding the *R* and *Y* relationship) and effect homogeneity.

It is possible to explore the existence of effect heterogeneity by comparing effect estimates across analytic approaches that target different estimands (i.e., ATT vs. SATT). The simulation results indicate that under the sequential ignorability assumption, the intact pairs and impute *R* approaches can accurately recover the SATT, while the other approaches are more sensitive to effect heterogeneity and confounding. In the empirical illustration, for example, the average treatment effect estimates were relatively stable across the analytic approaches. This suggests that effect heterogeneity is probably not a large concern regarding possible bias in treatment effect estimates that ignore TCM and between-strata effect heterogeneity. It does not say anything, however, about whether the average effect estimates are biased because of breakdowns in the sequential ignorability assumption. Here, researchers should probably use sensitivity analysis to explore the extent to which bias might arise from unobserved confounding factors (Imai, Keele, & Yamamoto, 2010).

In demonstrating TCM, we only explored a subset of possible analytic approaches one could use for treatment effect estimation and mediation analysis. Future research should examine whether other approaches are particularly robust to complications brought about by TCM. For example, more explicit Bayesian techniques for classifying units into principal strata (Page, 2012) may be more appropriate than the ad hoc impute *R* approach used in this paper. More importantly, we focused on the application of the principal stratification framework to TCM, but the appropriateness of this framework for TCM and mediation analysis more generally is not uniformly accepted (VanderWeele, 2012). In fact, weighting approaches have been applied to mediation analysis (Hong & Nomi, 2012) and to settings where, like TCM, selective attrition is a concern (Weuve et al., 2012). It would be valuable to see how these approaches perform under the different TCM conditions tested in this paper.

References

Allison, P. D. (2001). *Missing data*. SAGE.

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of Causal Effects Using Instrumental Variables. *Journal of the American Statistical Association*, *91*(434), 444–455.

Barnard, J., Frangakis, C. E., Hill, J. L., & Rubin, D. B. (2003). Principal Stratification Approach to Broken Randomized Experiments. *Journal of the American Statistical Association*, *98*(462), 299–323.

Bollen, K. A. (1987). Total, direct, and indirect effects in structural equation models. *Sociological methodology*, *17*(1), 37-69.

Frangakis, C. E., & Rubin, D. B. (2002). Principal Stratification in Causal Inference. *Biometrics*, *58*(1), 21–29.

Gallop, R., Small, D. S., Lin, J. Y., Elliott, M. R., Joffe, M., & Ten Have, T. R. (2009). Mediation analysis with principal stratification. *Statistics in Medicine*, *28*(7), 1108–1130.

Gleason, P., Clark, M., Tuttle, C. C., & Dwoyer, E. (2010). *The Evaluation of Charter School Impacts: Final Report. NCEE 2010-4029*. National Center for Education Evaluation and Regional Assistance.

Green, D. P., Ha, S. E., & Bullock, J. G. (2010). Enough Already about "Black Box" Experiments: Studying Mediation Is More Difficult than Most Scholars Suppose. *The ANNALS of the American Academy of Political and Social Science*, *628*(1), 200–208.

Holland, P. W. (1986). Statistics and Causal Inference. *Journal of the American Statistical Association*, *81*(396), 945–960.

Holland, P. W. (1988). Causal Inference, Path Analysis, and Recursive Structural Equations Models. *Sociological Methodology*, *18*, 449–484.

Heckman, J. J., & Hotz, V. J. (1989). Choosing Among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training. *Journal of the American Statistical Association*, *84*(408), 862–874.

Herman, J.L., Wang, J., Rickles, J., Hsu, V., Monroe, S., Leon, S., & Straubhaar, R. (2012). Evaluation of Green Dot's Locke Transformation Project: Findings for cohort 1 and 2 students (CRESST Report 815). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2011). MatchIt: Nonparametric Preprocessing for Parametric Causal Inference. *Journal of Statistical Software*, *42*(8), 1-28.

Hong, G., & Nomi, T. (2012). Weighting methods for assessing policy effects mediated by peer change. *Journal of Research on Educational Effectiveness*, *5*(3), 261-289.

Imai, K., Keele, L., & Yamamoto, T. (2010). Identification, Inference and Sensitivity Analysis for Causal Mediation Effects. *Statistical Science*, *25*(1), 51–71.

Jo, B. (2008). Causal inference in randomized experiments with mediational processes. *Psychological Methods*, *13*(4), 314–336.

Jo, B., Stuart, E. A., MacKinnon, D. P., & Vinokur, A. D. (2011). The Use of Propensity Scores in Mediation Analysis. *Multivariate Behavioral Research*, *46*(3), 425–452.

Judd, C. M., & Kenny, D. A. (1981). Process Analysis Estimating Mediation in Treatment Evaluations. *Evaluation Review*, *5*(5), 602–619.

Little, R. J., & Rubin, D. B. (2002). *Statistical Analysis with Missing Data, Second Edition* (2nd ed.). Wiley-Interscience.

McConnell, S., Stuart, E. A., & Devaney, B. (2008). The Truncation-by-Death Problem What To Do in an Experimental Evaluation When the Outcome Is Not Always Defined. *Evaluation Review*, *32*(2), 157–186.

Page, L. C. (2012). Principal Stratification as a Framework for Investigating Mediational Processes in Experimental Settings. *Journal of Research on Educational Effectiveness*, *5*(3), 215-244.

Pearl, J. (2009). *Causality: Models, Reasoning and Inference* (2nd ed.). Cambridge University Press.

Rickles, J., Wang, J., & Herman, J.L. (2013). Evaluation of Green Dot's Locke Transformation Project: Supplemental report on cohort 2 student outcomes (CRESST Report 825). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

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.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688–701.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581 –592.

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

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference* (2nd ed.). Boston, MA: Houghton-Mifflin.

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software. *Journal of Statistical Software*, *45*(3), 1-67.

VanderWeele, T. J. (2012). Comments: Should Principal Stratification Be Used to Study Mediational Processes?. *Journal of Research on Educational Effectiveness*, *5*(3), 245-249.

Weuve, J., Tchetgen Tchetgen, E. J., Glymour, M. M., Beck, T. L., Aggarwal, N. T., Wilson, R. S., … Mendes de Leon, C. F. (2012). Accounting for Bias Due to Selective Attrition. *Epidemiology*, *23*(1), 119–128.

Zhang, J. L., & Rubin, D. B. (2003). Estimation of Causal Effects Via Principal Stratification When Some Outcomes Are Truncated by "Death." *Journal of Educational and Behavioral Statistics*, *28*(4), 353–368.

Appendix A

Detailed Description of Analytic Approaches Used for the Simulation Study

Average treatment effects for both the simulation studies and the empirical illustration were estimated using the five approaches outlined in Table 3 based on a matched sample of treatment and control units and one ANCOVA approach based on the full (pre-match) sample. Details of each approach, as applied to the simulation study data, are presented below. The same general logic and steps were used for the empirical illustration, with minor details tailored to the specifics of the data and analysis. The notation used throughout follows from the theoretical discussion and simulation study. In addition to describing each approach, we include R code for executing each approach, as well as R code for constructing the Zhang & Rubin (2003) large sample bounds.

ANCOVA approach on full pre-matched sample. This approach uses a standard OLS regression model to adjust for confounding due to observed pretreatment factors (*X1* and *X2*):

$$Y_i = \beta_0 + \delta D_i + \beta_1 X1_i + \beta_2 X2_i + e_i \quad \text{(A.1)}$$

The analytic sample for this approach includes all units with non-missing outcome data.

```
lmx <- lm(Y~D+X1+X2, data=df)
summary(lmx)
```

Complete case analysis of original matched sample (1A). Besides the above ANCOVA approach, the other approaches are based on treatment and control units matched on their pretreatment covariates. For the simulation study, we used 1-to-1 nearest neighbor propensity score matching, where the estimated propensity score was based on the following logistic regression model:

$$\ln\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = \beta_0 + \beta_1 X1_i + \beta_2 X2_i. \quad \text{(A.2)}$$

We used the *matchit* package in R (Ho, Imai, King, & Stuart, 2011) to execute the matching. After matching treatment and control units on the estimated propensity score, the average treatment effect was estimated using the same regression model in Equation A.1.

```
## Conduct 1-to-1 Nearest Neighbor Match ##
df.m <- matchit(D ~ X1 + X2, data = df, method =
"nearest")

# Save Matched Data as data frame
```

```
df.md <- match.data(df.m, weights="PSCOREW",
distance="PSCORE")

# Merge matched data with main file
dfm<-merge(df,df.md[,c("ID","PSCOREW","PSCORE")],
by="ID",all.x=TRUE)
dfm$PSCOREW <- ifelse(is.na(dfm$PSCOREW),0,1)

## Estimate Average Treatment Effect ##
## with Matched Data
dfx <- dfm[dfm$PSCOREW>1,]
lmx <- lm(Y~D+X1+X2, data=dfx)
summary(lmx)
```

Multiple imputation of Y (1B). While the 1A approach is restricted to units with a non-missing outcome value, this approach uses the same matched sample but multiply imputes missing outcome values so the analysis is based on all the matched units. We used the *mice* package in R (van Buuren & Groothuis-Oudshoorn, 2011) to execute the multiple imputation and combine the effect estimates across the five multiply imputed data sets.

```
## Create 5 Imputed Data Sets ##
dfx<-dfm[dfm$PSCOREW>0,c("D","Y","X1","X2")]
imp<-mice(dfx,print=FALSE)

## Estimate Average Treatment Effect ##
## Pooling across the 5 data sets
lmx<-pool(lm.mids(Y~D+X1+X2,data=imp))
summary(lmx)
```

Rematching cases with observed Y (2A). This approach is the same as in 1A, but matching is only allowed among units with non-missing outcome data (*R*=1).

```
## REMATCH RESTRICTED TO RETAINED STUDENTS ##
df2 <- df[df$R==1,c("ID","X1","X2","D")]

## Conduct 1-to-1 Nearest Neighbor Match ##
df.m <- matchit(D ~ X1 + X2, data = df2, method =
"nearest")

# Save Matched Data as data frame
df.md <- match.data(df.m, weights="PSCOREW",
distance="PSCORE")

# Merge matched data with main file
```

```
dfm<-merge(df,df.md[,c("ID","PSCOREW","PSCORE")],
by="ID",all.x=TRUE)
dfm$PSCOREW <- ifelse(is.na(dfm$PSCOREW),0,1)

## Estimate Average Treatment Effect ##
## with Matched Data
dfx <- dfm[dfm$PSCOREW>1,]
lmx <- lm(Y~D+X1+X2, data=dfx)
summary(lmx)
```

Analysis of intact pairs (2B). With the original match for 1A, each treatment unit is matched to a control unit. These pairs can be used to approximate principal strata membership by comparing the observed mediator response for each unit in the pair. Given an interest in the SATT, we restrict the analysis to matched pairs with R=1 for both the treatment and control unit.

```
## GET MATCHED PAIRS FROM ORIGINAL MATCH ##
df.mp<-cbind(df[row.names(df.m$match.matrix),
c("ID","X1","X2")], df[df.m$match.matrix,
c("ID","X1","X2")])
names(df.mp)<- c("TID","TX1","TX2","CID","CX1","CX2")

ee.t<-dfm[,c("ID","Yc","R")]
names(ee.t)<-c("TID","TYc","TR")
ee.c<-ee.t
names(ee.c)<-c("CID","CYc","CR")

df.mp<-merge(df.mp,ee.t,by="TID",all.x=TRUE)
df.mp<-merge(df.mp,ee.c,by="CID",all.x=TRUE)

# Identify treatment student principal strata based
on matched control student outcome
df.mp$PSTRATA<-ifelse(df.mp$TR==1 & df.mp$CR==1,
"Y-Y","")
df.mp$PSTRATA<-ifelse(df.mp$TR==1 & df.mp$CR==0,
"Y-N",df.mp$PSTRATA)
df.mp$PSTRATA<-ifelse(df.mp$TR==0 & df.mp$CR==1,
"N-Y",df.mp$PSTRATA)
df.mp$PSTRATA<-ifelse(df.mp$TR==0 & df.mp$CR==0,
"N-N",df.mp$PSTRATA)

## Restrict Analysis to "Intact Pairs" ##
ipt<-as.data.frame(df.mp[df.mp$PSTRATA=="Y-Y",
"TID"]); names(ipt)<-c("ID")
ipc<-as.data.frame(df.mp[df.mp$PSTRATA=="Y-Y",
"CID"]); names(ipc)<-c("ID")
```

```
ip<-as.data.frame(rbind(ipt,ipc))
ip$intact<-1

dfx<-merge(dfm,ip,by="ID")

# Estimate Average Treatment Effect
dfx<-dfx[dfx$intact==1,]
lmx <- lm(Y~D+X1+X2, data=dfx)
summary(lmx)
```

Imputation of mediator counterfactual (2C). With the original match for 1A, this approach uses the available covariates to multiply impute the counterfactual mediator value for each unit to approximate principal strata membership. In our example, this means imputing whether the treatment units would have remained in school if assigned to control, and whether control units would have remained in school if assigned to treatment. For the imputation of *R*, we first use a logistic regression model to get parameter estimates of the relationship between the observed pretreatment covariates, including treatment indicator, and the probability of remaining in school (p_i):

$$\ln\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = \beta_0 + \beta_1 X1_i + \beta_2 X2_i + \beta_3 D_i. \qquad \text{(A.3)}$$

For each imputation, we then sample parameter values from a multivariate normal distribution with means centered on the model estimated parameter values and variances/covariances based on the model estimated standard errors. With the sampled parameters, $\hat{p}_i$ is calculated based on the observed pretreatment covariates and the alternative treatment value (i.e., 1-D_i in place of D_i). Each unit's counterfactual *R* value is then imputed based on whether a random uniform draw from 0 to 1 falls below $\hat{p}_i$ ($\hat{R}=1$) or above $\hat{p}_i$ ($\hat{R}=0$). Given an interest in the SATT, we then estimate the average treatment effect with the same ANCOVA model as in 1A, but only based on units in the matched sample with *R*=1 for both the observed condition and the counterfactual condition. We conducted 1,000 imputations of *R* and used standard multiple imputation combination rules to get the mean SATT and standard error across the 1,000 imputations.

```
## IMPUTATION OF COUNTERFACTUAL RETENTION VALUE ##
dfx <- dfm[dfm$PSCOREW==1,]

# Get model parameter estimates
XX <- as.matrix(dfx[,c("X1","X2")]) # covariates
R <- dfx$R # retention under actual treatment
D <- dfx$D # treatment indicator
```

```
Yobs <- dfx$Y # observed outcome

fm <- glm(R ~ XX + D,family=binomial("logit"))

# Impute R & estimate SATT 1,000 times
M<-1000 # number of imputations
ES <- NULL # storage for effect size estimates
EV <- NULL # storage for within imp error variances

for (m in 1:M) { # loop over imputations
  B.hat <- mvrnorm(1,coefficients(fm),vcov(fm))
  lo <- cbind(rep(1,length(D)),XX,(1-D))%*%B.hat
  p <- exp(lo)/(1+exp(lo))
  Rcf <- rep(0,length(D))
  Rcf[(runif(length(D),0,1) < p)] <- 1
  Y_RR <- Yobs[R==1 & Rcf==1]
  X_RR <- XX[(R==1 & Rcf==1),]
  D_RR <- D[R==1 & Rcf==1]
  fm2 <- lm(Y_RR ~ X_RR + D_RR)
  By <- coefficients(fm2)
  ES <- c(ES,By[length(By)])
  EV <- c(EV,vcov(fm2)[length(By),length(By)])
}

est <- cbind(mean(ES),sqrt(mean(EV)+(1+1/M)*var(ES)))
```

Large sample bounds. For the empirical illustration, we also calculated the large sample bounds following Zhang & Rubin (2003). The bounds are calculated under different assumptions about the principal strata.

```
## Large-sample bounds (Zhang and Rubin, 2003) ##
dfx <- dfm[dfm$PSCOREW==1,]
R <- dfx$R # retention under actual treatment
D <- dfx$D # treatment indicator
Yobs <- dfx$Y # observed outcome

# observed groups
CG <- Yobs[(D==0 & R==1)]
TG <- Yobs[(D==1 & R==1)]
P_CG <- length(CG)/length(Yobs[D==0])
P_TG <- length(TG)/length(Yobs[D==1])
minpi_DG <- max(0,P_CG-P_TG)
maxpi_DG <- min(P_CG,(1-P_TG))
minESnoA <- NULL # ES lower bound no assumptions
maxESnoA <- NULL # ES upper bound no assumptions
minESA2 <- NULL # ES lower bound ranked ave score
```

```
maxESA2 <- NULL # ES upper bound /ranked ave score

for (pi_DG in seq(minpi_DG,maxpi_DG,0.001)) {
 minESnoA <- c(minESnoA,
  mean(sort(TG,decreasing=FALSE)[1:min(((P_CG/P_TG-
   pi_DG/P_TG)*length(TG)),length(TG))]) -
  mean(sort(CG,decreasing=TRUE)[1:min(((1-
   (pi_DG/P_CG))*length(CG)),length(CG))]))

 maxESnoA <- c(maxESnoA,
  mean(sort(TG,decreasing=TRUE)[1:min(((P_CG/P_TG-
   pi_DG/P_TG)*length(TG)),length(TG))]) -
  mean(sort(CG,decreasing=FALSE)[1:min(((1-
   (pi_DG/P_CG))*length(CG)),length(CG))]))

 minESA2 <- c(minESA2,
  mean(sort(CG,decreasing=TRUE)[1:min(((1-
   pi_DG/P_CG)*length(CG)),length(CG))]))

 maxESA2 <- c(maxESA2,
  mean(sort(TG,decreasing=TRUE)[1:min((max(1,
  (P_CG/P_TG-pi_DG/P_CG)*length(TG))),length(TG))])-
  mean(CG))

 #no assmputions ("noA")
  minESnoA <- min(minESnoA)
  maxESnoA <- max(maxESnoA)

 #monotonicity assumption ("A1") - pi_DG=0
  minESA1 <- mean(sort(TG,
   decreasing=FALSE)[1:min(((P_CG/P_TG)*length(TG)),
   length(TG))]) - mean(CG)

  maxESA1 <- mean(sort(TG,
   decreasing=TRUE)[1:min(((P_CG/P_TG)*length(TG)),
   length(TG))]) - mean(CG)

 #ranked average score assumption ("A2")
  minESA2 <- mean(TG)-max(minESA2)
  maxESA2 <- max(maxESA2)

 # monotonicity & ranked average score ("A1A2")
  minESA1A2 <- mean(TG)-mean(CG)
  maxESA1A2 <- mean(sort(TG,
   decreasing=TRUE)[1:min(((P_CG/P_TG)*length(TG)),
   length(TG))])-mean(CG)
```

```
  B <- matrix(c(minESnoA,maxESnoA,minESA1,maxESA1,
   minESA2,maxESA2,minESA1A2,maxESA1A2),
   4,2,byrow=TRUE)
  rownames(B) <- c("none","A1","A2","A1+A2")
  colnames(B) <- c("lower","upper")
}
```

Appendix B

Summary of Simulation Study Results

Table B1

Simulation study 1 results, by condition and estimation method

		Proportion of Encouragers to Always Stayers								
		Small			Medium			Large		
U-R Effect	Method	Bias	RMSE	Type I	Bias	RMSE	Type I	Bias	RMSE	Type I
None	ANCOVA	-0.001	0.031	0.041	0.002	0.033	0.051	-0.001	0.037	0.043
	Main Match	-0.001	0.034	0.042	0.002	0.038	0.057	-0.002	0.045	0.050
	Re-Match	-0.001	0.034	0.044	0.002	0.036	0.060	-0.001	0.038	0.047
	Intact Pairs	-0.001	0.037	0.039	0.001	0.044	0.050	0.000	0.060	0.045
	Impute Y	0.000	0.035	0.048	0.002	0.039	0.059	-0.002	0.047	0.064
	Impute R	-0.001	0.034	0.016	0.002	0.039	0.012	-0.002	0.046	0.004
Moderate	ANCOVA	-0.026	0.041	0.131	-0.046	0.056	0.262	-0.090	0.098	0.669
	Main Match	-0.024	0.042	0.105	-0.041	0.055	0.187	-0.081	0.093	0.425
	Re-Match	-0.024	0.043	0.111	-0.041	0.053	0.211	-0.086	0.095	0.583
	Intact Pairs	-0.021	0.044	0.081	-0.037	0.056	0.134	-0.076	0.096	0.256
	Impute Y	-0.023	0.042	0.100	-0.041	0.055	0.183	-0.079	0.093	0.426
	Impute R	-0.021	0.041	0.051	-0.037	0.052	0.045	-0.073	0.087	0.086
Large	ANCOVA	-0.056	0.064	0.392	-0.093	0.099	0.784	-0.179	0.183	0.998
	Main Match	-0.051	0.062	0.294	-0.085	0.093	0.600	-0.162	0.168	0.941
	Re-Match	-0.049	0.060	0.270	-0.085	0.092	0.651	-0.171	0.175	0.990
	Intact Pairs	-0.045	0.059	0.211	-0.078	0.089	0.424	-0.150	0.161	0.703
	Impute Y	-0.049	0.061	0.278	-0.084	0.092	0.567	-0.160	0.166	0.898
	Impute R	-0.046	0.058	0.152	-0.077	0.086	0.298	-0.147	0.154	0.563

Table B2

Simulation study 2 results, by condition and estimation method

Effect Heterogeneity	Method	Proportion of Encouragers to Always Stayers: Small Bias	Small RMSE	Small Coverage	Medium Bias	Medium RMSE	Medium Coverage	Large Bias	Large RMSE	Large Coverage
Small	ANCOVA	-0.053	0.062	0.634	-0.065	0.073	0.520	-0.103	0.110	0.241
	Main Match	-0.021	0.041	0.926	-0.026	0.047	0.898	-0.049	0.068	0.828
	Re-Match	-0.018	0.040	0.937	-0.034	0.050	0.836	-0.079	0.089	0.506
	Intact Pairs	0.005	0.038	0.960	0.003	0.046	0.938	-0.014	0.063	0.944
	Impute Y	-0.021	0.041	0.930	-0.026	0.047	0.893	-0.049	0.069	0.803
	Impute R	0.001	0.035	0.981	0.001	0.040	0.987	-0.013	0.050	0.992
Medium	ANCOVA	-0.102	0.108	0.178	-0.133	0.138	0.050	-0.204	0.208	0.002
	Main Match	-0.038	0.055	0.818	-0.055	0.068	0.745	-0.094	0.108	0.512
	Re-Match	-0.034	0.053	0.846	-0.071	0.081	0.541	-0.156	0.165	0.109
	Intact Pairs	0.015	0.047	0.924	0.006	0.048	0.949	-0.028	0.073	0.907
	Impute Y	-0.042	0.058	0.805	-0.059	0.072	0.729	-0.096	0.110	0.498
	Impute R	0.005	0.041	0.973	-0.002	0.042	0.991	-0.023	0.060	0.985
Large	ANCOVA	-0.208	0.213	0.003	-0.269	0.273	0.000	-0.404	0.407	0.000
	Main Match	-0.082	0.096	0.573	-0.115	0.127	0.363	-0.183	0.195	0.132
	Re-Match	-0.072	0.087	0.632	-0.146	0.154	0.127	-0.308	0.318	0.009
	Intact Pairs	0.026	0.061	0.901	0.007	0.061	0.927	-0.048	0.096	0.882
	Impute Y	-0.101	0.112	0.435	-0.132	0.142	0.244	-0.194	0.205	0.126
	Impute R	0.005	0.051	0.965	-0.008	0.057	0.978	-0.039	0.082	0.972