2-3-2017

IT'S ALL ABOUT BALANCE: PROPENSITY SCORE MATCHING IN THE CONTEXT OF COMPLEX SURVEY DATA

David Lenis
Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, dlenis@jhsph.edu

Trang Q. ;Nguyen
Department of Mental Health, Johns Hopkins Bloomberg School of Public Health

Nian Dong
Department of Educational, School and Counseling Psychology, University of Missouri

Elizabeth A. Stuart
Departments of Mental Health, Biostatistics and Health Policy and Management, Johns Hopkins Bloomberg School of Public Health

It's all about balance: propensity score matching in the context of complex survey data

David Lenis[1], Trang Q. Nguyen[2], Nianbo Dong[3], and Elizabeth A. Stuart[4]

[1]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21205, USA
[2]Departments of Mental Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21205, USA
[3]Department of Educational, School and Counseling Psychology, University of Missouri, Columbia, MO 65211, USA
[4]Departments of Mental Health, Biostatistics, and Health Policy and Management, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21205, USA

Abstract

Many research studies aim to draw causal inferences using data from large, nationally representative survey samples, and many of these studies use propensity score matching to make those causal inferences as rigorous as possible given the non-experimental nature of the data. However, very few applied studies are careful about incorporating the survey design with the propensity score analysis, which may mean that the results don't generate population inferences. This may be because few methodological studies examine how to best combine these methods. Furthermore, even fewer of the methodological studies incorporate different non-response mechanisms in their analysis. This study examines methods for how to handle survey weights in propensity score matching analyses of survey data, under different non-response mechanisms. Based on the results from Monte Carlo simulations implemented on synthetic data as well as a data based application we developed suggestions regarding the implementation of propensity score methods to make causal inferences relevant to the target population of a sample survey. Our main conclusions are: (1) whether the survey weights are incorporated in the estimation of the propensity score does not impact estimation of the population treatment effect, as long as good population balance is achieved across confounders, (2) survey weights must be taken into account in the outcome analysis and (3) transfer of survey weights (i.e., matched comparison units are assigned the sampling weight of the treated unit they have been matched to) can be beneficial under certain non-response mechanisms.

1 Introduction

1.1 Background

Much education research aims to make statements regarding the effects of interventions, such as a new reading program or being held back in kindergarten, on students' outcomes. Although randomized trials are seen as the gold standard for estimating causal effects, they are often infeasible, especially if interest is in the effects for a broad, representative sample of students, such as all kindergarteners across the U.S. Answering causal questions and obtaining population causal effect estimates that generalize nationally often requires existing non-experimental data, as well as statistical methods to ensure the inferences are as accurate as possible. Two tools that are available to help with these answers are large-scale nationally-representative datasets and propensity score methods, which help control for differences between those who do and do not receive some intervention or exposure of interest.

Large scale, complex survey designs are widely used in educational research. A paradigmatic example is the Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K). Surveys such as this usually have a well-defined target population and sampling framework. In such large-scale surveys, the sampling framework may be complicated, and the sampling probabilities (and resulting survey weights) vary depending on the sampling of sub-populations (these survey weights may then also be adjusted to account for non-response or to post-stratify to known population totals). When attempting to make inferences to the target population, the survey design (e.g., survey weights and design elements) should be correctly used in data analysis; otherwise, the parameter estimates may not be relevant to the original target population of the survey (see, e.g., Hansen et al. [1983], Korn and Graubard [1995a], Korn and Graubard [1995b] and Little [2003]).

In addition to descriptive and correlational analyses, it is often of interest to draw causal inferences using data from complex surveys. One of the main contributions of the field of causal inference is that has provided a number of tools and a clear conceptual framework in which causal effects can be estimated. Furthermore, the development of causal methods has had an important impact in the design of clinical trials, but also, the causal inference framework has been extended to non-experimental studies where formal randomization is not possible. One milestone in the causal inference framework was the definition and implementation of propensity score based methods. The propensity score is defined as the probability of receiving treatment given a set of observed covariates and it was first introduced by Rosenbaum and Rubin [1983]. Since then, a wide range of propensity score based methods have been developed to estimate treatment effects in non-experimental studies, including matching, weighting, and subclassification (Stuart [2010]). These methods are often implemented to help make causal statements when a randomized experiment is infeasible.

In particular, matching estimators have been widely used in the context of non-experimental studies. They help reduce bias in the estimation of causal effects (seeRubin [1973], and are intuitive and relatively easy to implement. One approach is to match observations (i.e., comparison to treated units) on a set of observed covariates. The main goal of this matching

approach is to generate a new sample (i.e., the matched sample), such that for every treated unit there is (at least) one comparison unit with similar values of observed covariates. The outcome of interest is then compared between the matched treated and matched comparison subjects to estimate the causal effect. One main disadvantage of this procedure resides in the fact that as the number of variables on which units are matched on increases, the chances of finding matched pairs with similar observed characteristics decreases exponentially (this issue is often referenced as "the curse of dimensionality"). Thus, matching directly on a set of covariates is only feasible in large samples and/or if a small set of covariates are used in the matching procedure. Nevertheless, one of the main contributions of Rosenbaum and Rubin [1983] was to show that a similar (or balanced) distribution of the observed characteristics can also be achieved when the matching procedure is based on the propensity score instead of the entire set of observed covariates.

Nonetheless, standard application of propensity score matching methods doesn't give guidance on how to incorporate survey weights and it is conceptually unclear how to do so (i.e., whether the survey weights need to be incorporated in the estimation of the propensity score model and whether they should be used in the outcome analysis). As results, often researchers using propensity score matching methods don't incorporate the complex survey design. This paper aims to provide guidance, to ensure that the results apply to the target population.

1.2 Previous Research in this Area

There has been extensive work in each of the two areas to be investigated in this article (complex surveys and propensity scores), but only limited work on how to combine them.

Although propensity methods have been developed under the assumption of a simple random sample (SRS), this sampling scheme is hardly ever used. On one hand a SRS requires that the survey team has access to a full list of the units that compose the population of interest, which in most situation is not available. On the other hand, SRS may lead to skewed representation of some population features and therefore additional sampling techniques may be required. To guarantee representation of all relevant features in the population, complex survey techniques may be implemented. Some of such techniques divide the population in groups such strata and clusters and sample from them. Both strata and cluster divide the population into mutually exclusive and exhaustive groups. The main difference between stratum and cluster resides on the fact that every strata is sampled (i.e., represented in the final sample) but not every cluster is. The sampling design in addition to certain adjustments (e.g., non-response or post-stratification) define the survey weights that should be used to scale the sample back to the population.

There is a generalized consensus that ignoring survey weights leads mainly to external validity bias, where inferences about the sampled population are based on a unrepresentative analytic sample. Thus, survey weights and the sampling design should be incorporated in the estimation process. It has been widely documented how to incorporate the survey weights in the estimation of means, totals and ratios (see Cochran [1977] and Groves et al. [2009]), nonetheless there is still some controversy on how to incorporate survey weights in more

complex statistical methods (see Gelman [2007]), and propensity score methods are not an exemption.

Propensity score methods, include two key stages: (1) estimating propensity scores, and (2) using those propensity scores (e.g., through matching, weighting, or subclassification) to estimate causal effects. And in each of these two stages it is unclear how (or if) survey weights should be included.

When estimating the propensity score, Brunell and DiNardo [2004] and Heckman and Todd [2009] argue that misspecified estimates of propensity scores (e.g., using unweighted propensity score models) do not cause problems in causal inference because "the odds ratio of the propensity score estimated using misspecified weights is monotonically related to the odds ratio of the true propensity scores" (Heckman and Todd [2009] p.3), "and therefore does not change the relative weighting of the data" (Brunell and DiNardo [2004], p.32). However, it is not clear how broadly applicable these results are and the theoretical results in these papers have not been empirically validated. Others like Ridgeway et al. [2015] argue (in the context of propensity score weighting) that survey weights should be incorporated in the estimation of the propensity score, and failure to do so may lead to inconsistent estimators. Work by Little [2003] discussed both weighted and unweighted models in the context of non-response adjustment (which uses similar methods to propensity scores in non-experimental data). Little and Vartivarian [2003] argued that although the weighted response rate is an unbiased estimate of the population parameter, this does not ensure unbiased estimates of the variables of interest. Grau et al. [2006] and Potter et al. [2006] empirically examined Little and Vartivarian's assertion by evaluating the use of weighted and unweighted logistic regression models to estimate propensity scores for survey nonresponse adjustment. Grau et al. [2006] and Potter et al. [2006] did not find significant differences in the parameter estimates between the two options. Thus, there has been some initial investigation of the use of weighted versus unweighted models for estimating the propensity score, at least in the survey nonresponse context, but no clear consensus. Nevertheless, little work has addressed how to combine propensity scores and complex survey data in a causal inference framework. Until recently, most of the existing work (as limited as it is) has been done in the context of subclassification (see Hong and Raudenbush [2005]) or weighting (see Ridgeway et al. [2015]), but little work has investigated the performance for propensity score matching in particular (see Austin et al. [2016]).

The use of the propensity score, unsurprisingly, also raises questions regarding how the survey weights should used. For example, after implementing a propensity score matching procedure, do survey weights need to be incorporated to asses the balance of the covariates? or in the context of propensity score weighting: how should the final weights be constructed?. Authors like Ridgeway et al. [2015] argue that survey weights should not only be used in the estimation of the propensity score, but they should be combined with the propensity score weights when executing the outcome analysis.

To be more specific, with respect to survey weights in particular, our review of the education literature that uses propensity score methods to estimate causal effects using complex surveys has found at least five approaches to handling survey weights:

- ignoring survey weights entirely in both propensity score estimation and use (e.g., Hallberg et al. [2011], Hong and Raudenbush [2005], Morgan et al. [2008]),
- using survey weights as a covariate when estimating propensity scores (Stage 1; e.g., Korenman et al. [2013]) and in the matching (Stage 2; e.g., Rubin [2001]),
- a complex method that uses matching on the propensity scores (but no discussion of the weights in estimating those propensity scores) and then makes the sum of the sample weights equal for the matched treated and matched comparison groups, while also keeping the modified weights proportional to the original weights (Reardon et al. [2009]),
- estimating the propensity score using an unweighted model (Stage 1), and then in Stage 2 stratifying on the propensity score and estimating causal effects within each subclass using a survey-weighted regression of the outcome (e.g., Hahs-Vaughn and Onwuegbuzie [2006], Zanutto et al. [2005]; Zanutto [2006]), and
- multiplying the original survey weights by the propensity score weights and using those combined weights in the outcome analysis (Stage 2), with no discussion of whether the propensity score estimation (Stage 1) accounted for the complex survey design ('Zanutto et al. [2005], Schonlau et al. [2004]).

There is thus a broad array of approaches used in the literature, and until recently there was almost no methodological work on the best ways to use propensity scores with complex surveys, with the exception of Zanutto [2006], Ridgeway et al. [2015] and Austin et al. [2016]. Austin et al. [2016] explores the implementation of propensity score matching in the context of complex survey data. The authors explore three alternative ways to incorporate the survey weights in the estimation of the propensity score (i.e., using the weights as a covariate in the propensity score model, using them as weights in a weighted propensity score model, and not using the survey weights at all in the estimation of the propensity score). They also explore whether or not weights should be transferred between treated and the comparison units that they have been matched to. In general terms, their simulation results indicate that there is not a specific way to incorporate the survey weights that results in a better performance (measured in terms of bias, mean squared error and coverage of a 95% confidence interval).

In this article we extend the scope of the analysis of previous research to incorporate different non-response mechanisms. Non-response is a phenomenon by which is not possible to collect the desire information for some observations and tends to be rule rather the exception in complex survey data. To our knowledge, previous research in this area has not addressed the consequences associated with this common phenomenon. The reason why this extension is relevant is two-fold: (1) it will allow to evaluate the performance of different matching estimators in a more realistic fashion and (2) it will allow to identify if features of the non-response mechanism can impact the performance of the matching estimators. Furthermore, this extension allows us to incorporate in our analysis the fact that most of survey weights include adjustments for non-response. In this way, the main goal of this article is to identify

ways that the survey weights should be incorporated when using propensity score matching to estimate causal effects, under different non-response mechanisms.

The rest of this article is organized as follows: in Section 2 we discuss the definitions and assumptions involved in the estimation of the average causal effect and also how survey weights should be incorporated in the estimation procedure. Section 3 describes the simulation study implemented and summarizes our main findings. Section 4 compares the performance of the different estimation procedure in an application using the Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ESCL-K) and in Section 5 we present our main conclusions and discussion.

2 Definitions, Assumptions, Propensity Score and Survey Weights

2.1 Definitions and Assumptions

The Causal Inference Framework

Traditionally, causal treatment effects are defined in the context of the Rubin Causal Model (RCM) (see Rubin [1974]). In the RCM an individual treatment effect, associated with a binary treatment assignment T (that takes the value 1 if the sample unit receives the treatment of interest and 0 otherwise), is defined in terms of potential outcomes. For each unit i, $Y_i(t)$ with $t = 0, 1$ represents the outcome that would have been observed if unit i received the treatment t. Thus, the treatment effect associated with unit i is defined as $Y_i(1) - Y_i(0)$. Nevertheless, notice that for any given unit i the pair $(Y_i(0), Y_i(1))$ is not observable, we only observe one of the two potential outcome. Explicitly the observed outcome, Y_i, is defined as:

$$Y_i = Y_i(1) \times T_i + Y_i(0) \times (1 - T_i) \tag{1}$$

Equation (1) is referred as the "consistency of the observed outcome assumption" (see Hernán and Robins [2017]). Given that the unit level treatment effects cannot be estimated, we are often interested in estimating average treatment effects. At the population level, the most commonly defined average effects are: (1) the population average treatment effect (PATE) and (2) the population average treatment effect on the treated (PATT).

The PATE is defined as average effect across the population. Explicitly:

$$PATE = E\left[Y_i(1) - Y_i(0)\right] \tag{2}$$

Under randomization of the treatment, units in the treated group and the units in the control group have similar distributions of covariates (observed and unobserved) and potential outcomes. In this way, the average outcome computed among the units in the treated groups serves as a good counterfactual for the average outcome computed among the units

in the control group. The differences between these two averages is an estimator of the population average treatment effect (PATE).

Another treatment effect that traditionally is considered in the causal inference literature is the average treatment effect on the treated (ATT). The ATT is defined as the average gain, in the population, from the treatment computed among those units who were actually treated. In symbols:

$$PATT = E\left[Y_i(1) - Y_i(0)|T_i = 1\right] \tag{3}$$

When the treatment is randomized and the treatment effect is homogeneous (the treatment effect is the same for all units in the sample), it holds that the PATE is equal to the PATT. Nevertheless, when the treatment effect is not homogeneous the PATT and the PATE can be quite different.

When randomization is not feasible, additional assumptions are required to identify and estimate treatment effects. In particular, a crucial assumption in the estimation of treatment effects is the one referred to as "ignorability" (see Hernán and Robins [2017]). To further describe the implications of this assumption, define for all i, $\mathbf{X}_i$ as $q-$dimensional vector of covariates, i.e., $\mathbf{X}_i = (X_{1,i}, \ldots, X_{q,i})$. Ignorability assumes that $\mathbf{X}$ contains all possible confounders. In other words, given the set of observed covariates $\mathbf{X}$, the treatment assignment is independent of the potential outcomes. The ignorability assumption means that the treatment assignment can be assumed to be random, conditionally on observable characteristics of the units in the sample. This implies that:

$$(Y_i(0); Y_i(1)) \perp\!\!\!\perp T_i|\mathbf{X}_i \tag{4}$$

Another key assumption of the RCM is the Stable Unit Treatment Value Assumption (or **SUTVA**). The implication of this assumption is twofold: (1) the treatment assignment of any unit does not affect the potential outcomes of other units (often referred to as non-interference) and (2) there is only one version of the treatment, this implies that the treatment is comparable across units (see Hernán and Robins [2017]).

PATT vs. SATT

Ideally, we would like to estimate population causal effects but the instances where data on the full target population is available are quite unusual. In reality, causal effects are estimated using data from a sample. For example, the sample ATE (or SATE) represents the difference in average outcomes if everyone in the survey sample received the treatment versus everyone in the survey sample receiving the control condition. When does the SATE and the sample ATT (or SATT) correctly estimates the PATE and the PATT? The answer to this question depends on two key factors: (1) the sampling design and (2) the non-response mechanism.

The SATE (SATT) will correctly estimate the PATE (PATT) only when the sample distribution of the relevant variables is similar to its population counterpart. One sampling design that guarantees this is a simple random sample (SRS), but this kind of sampling

technique is hardly ever used. In general, most of available samples are the result of complex sampling designs. Therefore, unless survey weights are used to weight the sample back to the population, using sample information to estimate a treatment effect will result in a consistent estimator for the SATE (SATT) but not for the PATE (PATT).

In addition to the sampling design, the nature of the non-response mechanism, can potentially impact the estimation of the PATE (PATT). Non-response is a common phenomenon in the context of survey data and occurs when it is not possible to collect the desired information for some of the primary sampling units. In other words, for some units, some information is missing. Traditionally, missing data mechanisms are grouped in three categories: (1) Missing Completely at Random (MCAR), (2) Missing and Random (MAR) and (3) Missing not at Random (MNAR). Under a MCAR mechanism, the probability that one observation will have missing information, is completely random. In other words, there is no relationship between the propensity of the data to be missing and the values of the variables in the data set. When the non-response follows a MAR mechanism, the propensity of the data to be missing is random, conditional on the set of observed variables. In other words the observed values of the available data, can predict the probability of one observation to have missing information. Finally when the non-response is MNAR, the probability of having missing information depends on unobserved variables. That is, even after accounting for the observed variables available in the data, the propensity of the data to be missing is not random. Notice that even if the sampling design is such that the SATE (SATT) can be use to estimate the PATE (PATT), if the non-response mechanism is either MAR or MNAR the SATE (SATT) and the PATE (PATT) can be drastically different. Since most of survey weights include a non-response adjustment, failure to include them in the estimation procedure may result in misleading estimation results.

2.2 Survey Weights and the Propensity Score

In this section we formalize the the non-response mechanisms and the propensity score model. Consider a binary indicator S_i that takes the value 1 if the i^{th} unit has been selected into the survey sample and 0 otherwise. Additionally consider a response indicator, R_i, which takes the value 1 if unit i responds to the survey. We assume that at the population level each $O_i = (\mathbf{X_i}, T_i, Y_i, S_i, R_i)$ is independent and identically distributed with a joint density function f with $f : \mathbb{R}^{q+1} \times \{0,1\}^3 \rightarrow \mathbb{R}^+$. We represent the marginal distribution for a subset of covariates $\mathbf{Z}$ (i.e., $\mathbf{Z} \subset \mathbf{X}$) with $f_{\mathbf{Z}}$. We assume that the survey sample has finite size of $n = \sum_{i=1}^{N} SR_i$, where N represents the population size, and for every $i = 1, 2, ..., n$ $SR_i = S_i \times R_i$. Notice that SR_i constitutes a indicator variable that takes the value 1 if the sample unit i is selected into the survey and such unit responds. We consider the case where the probability for unit i of being observed in the sample (i.e., $SR_i = 1$) is function of a $q-$dimensional vector of covariates $\mathbf{X}$ and potentially the treatment indicator (T). Explicitly we assume that

$$p = f_{SR|(\mathbf{X},Y)}\left(SR = 1|\mathbf{X}, T\right) \tag{5}$$

where $f_{SR|(\mathbf{X},T)} : \mathbb{R}^{q+1} \to (0,1)$ and p is represents the probability for unit i of being in the final sample. We assume that $p \in (0,1)$ (i.e., there is not a set of values of the covariates X and T for which the probability of being in the sample is exactly 1 or exactly 0). Furthermore we assume that the survey weights, ω, are equal to the inverse of the probability of being observed in the sample, formally:

$$\omega = \frac{1}{p} = \frac{1}{f_{SR|(\mathbf{X},T)}\left(SR = 1|\mathbf{X}, T\right)} \tag{6}$$

Notice that this definition allows for a non-response rate different from 0 and different non-response mechanisms. To see this, consider the case where S and R are independent conditional on $(\mathbf{X}, T)$. Then it holds that $f_{SR|(\mathbf{X},T)}\left(SR = 1|\mathbf{X}, T\right)$ it's equal to $f_{S|(\mathbf{X},T)}\left(S = 1|\mathbf{X}, T\right)$ times $f_{R|(\mathbf{X},T)}\left(R = 1|\mathbf{X}, T\right)$, this last term models the non-response mechanism. Notice that if $f_{R|(\mathbf{X},T)}\left(R = 1|\mathbf{X}, T\right) = 1$ for all $(\mathbf{x}, t)$ in $(\mathbf{X}, T)$ then the non-response rate is 0. If $f_{R|(\mathbf{X},T)}\left(R = 1|\mathbf{X}, T\right) = f_R\left(R = 1\right)$ the non-respond mechanism is MCAR. Finally, the non response could be MAR and NMAR depending on whether the all the elements in $(\mathbf{X}, T)$ are observed. If every element in $(\mathbf{X}, T)$ is available to estimate the probability of non-response, then the non-response mechanism is MAR, otherwise the non-response process is NMAR. In this way, ω (the final observed sampling weight) is a combination of the survey weights associated with the sampling design itself but also incorporates corrections associated with non-response.

We define the propensity score as:

$$\pi = f_{T|X}\left(T = 1|\mathbf{X}\right) \tag{7}$$

which represents the probability of receiving treatment, conditional on $\mathbf{X}$, with $f_{T|\mathbf{X}} : \mathbb{R}^q \to (0,1)$. Notice that π represents the probability of receiving treatment computed at a population level, furthermore the same set of regressors are used in calculating p^s and π. This implies that all the information used in the construction of the survey weights is available in the survey sample. Notice that, in principle, π can be estimated in two ways: (1) incorporating the survey weights in the estimation procedure or by (2) computing π^S (the probability of receiving treatment computed among the units that compose the survey sample) with $\pi^S = f_{T|(\mathbf{X},RS)}\left(T = 1|\mathbf{X}, RS = 1\right)$, with $f_{T|\mathbf{X},RS} : \mathbb{R}^{q+1} \to (0,1)$. Thus, π^S can be estimated using the sample units without incorporating the survey weights in the estimation procedure.

2.3 Survey weights after matching

Traditionally, matching procedures are implemented to estimate the causal effects among the units who received treatment, work by Abadie and Imbens [2006] and Abadie and Imbens [2016] show that the average causal effects can also be estimated using matching procedures. Throughout this article, we focus on estimating causal effects among the treated.

As we stated previously, one of the main issues regarding the computation of causal effects using a propensity score matching procedure in the context of complex survey data, is

how, when or even if the survey weights should be incorporated. In this section we argue that survey weights may not need to be incorporated in the estimation of the propensity score model (stage 1), and show that the weights of the treated units should be transferred to the comparison unit to which they have been matched to, before estimating the outcome model. To see this, consider the following strategy: in a first step we implemented a matching procedure using the predicted propensity score (either the $\widehat{\pi^S}$ or $\widehat{\pi}$ can be used in the matching procedure, furthermore the survey weights could even be incorporated as an additional covariate). We assume that k comparison units were matched without replacement to each treated observation. Now, in order to identify the weights for the treated (w^t) and comparison units (w^c) to use in the outcome analysis, we note that under a successful implementation of the matching procedure it should hold that for every $\mathbf{x}$ in $\mathbf{X}$, the following equations hold:

$$f_{\mathbf{X}|T}(\mathbf{X}=\mathbf{x}|T=1) = w^c(\mathbf{x}) \times f_{\mathbf{x}|(T,M)}(\mathbf{X}=\mathbf{x}|T=1,M=1) \quad (8)$$
$$f_{\mathbf{X}|T}(\mathbf{X}=\mathbf{x}|T=1) = w^t(\mathbf{x}) \times f_{\mathbf{X}|(T,M)}(\mathbf{X}=\mathbf{x}|T=0,M=1) \quad (9)$$

In other words, after weighting, we want the distribution of the covariates among treated and comparison units in the matched sample ($M=1$), to be the similar to the distribution of the covariates among the treated at the population level (recall that we are interested in estimating the PATT).

From (9) we obtain that

$$\begin{aligned} w^t(\mathbf{x}) &= \frac{f_{\mathbf{X}|T}(\mathbf{X}=\mathbf{x}|T=1)}{f_{\mathbf{X}|(T,M)}(\mathbf{X}=\mathbf{x}|T=1,M=1)} \\ &= \frac{f_{(\mathbf{X},T)}(\mathbf{X}=\mathbf{x},T=1)}{f_T(T=1)} \times \frac{f_{(T,M)}(T=1,M=1)}{f_{(\mathbf{X},T,M)}(\mathbf{X}=\mathbf{x},T=1,M=1)} \\ &= \frac{f_{M|T}(M=1|T)}{f_{M|(\mathbf{X},T)}(M=1|\mathbf{X}=\mathbf{x},T=1)} \end{aligned} \quad (10)$$

Now, without trimming the treated units in the survey sample we have that (in the context of propensity score matching, often we find situations where the overlap of the predicted propensity score between treated and comparison units is poor, therefore is not always possible in to match every treated unit to k comparison observations. A common practice in this situations, is to drop such treated units from the analysis. See Frölich [2004]):

$$\begin{aligned} f_{M|T}(M=1|T=1) &= f_{SR|T}(SR=1|T=1) \\ f_{M|(\mathbf{X},T)}(M=1|\mathbf{X}=\mathbf{x},T=1) &= f_{SR|(\mathbf{X},T)}(SR=1|\mathbf{X}=\mathbf{x},T=1) \end{aligned}$$

Thus (10) can be expressed as:

$$
\begin{aligned}
w^t(\mathbf{x}) &= \frac{1}{f_{SR|(\mathbf{X},T)}(SR=1|\mathbf{X}=\mathbf{x},T=1)} \\
&= \omega
\end{aligned} \tag{11}
$$

Therefore we can conclude that units in the treated groups should be weighted using the survey weights assigned by the survey design.

Combining (8) , (9) and (11) allows us find the expression for the weights of the comparison units:

$$
\begin{aligned}
w^c(\mathbf{x}) &= \omega^t(\mathbf{x}) \times \frac{f_{\mathbf{X}|(T,M)}(\mathbf{X}=\mathbf{x}|T=1,M=1)}{f_{\mathbf{X}|(T,M)}(\mathbf{X}=\mathbf{x}|T=0,M=1)} \\
&= \omega^t(\mathbf{x}) \times \frac{f_{(X,T,M)}(\mathbf{X}=\mathbf{x},T=1,M=1)}{f_{(T,M)}(T=1,M=1)} \times \frac{f_{(T,M)}(T=0,M=1)}{f_{(X,T,M)}(\mathbf{X}=\mathbf{x},T=0,M=1)} \\
&= \omega^t(\mathbf{x}) \times \frac{f_{(T|\mathbf{X},M)}(T=1|\mathbf{X}=\mathbf{x},M=1)}{f_{(T|\mathbf{X},M)}(T=0|\mathbf{X}=\mathbf{x},M=1)} \times \frac{f_{(T,M)}(T=0,M=1)}{f_{(T,M)}(T=1,M=1)} \\
&= \omega^t(\mathbf{x}) \times \frac{f_{(T|\mathbf{X},M)}(T=1|\mathbf{X}=\mathbf{x},M=1)}{1-f_{(T|\mathbf{X},M)}(T=1|\mathbf{X}=\mathbf{x},M=1)} \times \frac{f_{(T|M)}(T=0|M=1)}{f_{(T|M)}(T=1|M=1)}
\end{aligned} \tag{12}
$$

Where $f_{T|(\mathbf{X},M)}(T=1|\mathbf{X}=\mathbf{x},M=1)$ is the value of the propensity score computed among the matched observations. Since we implemented a $k:1$ matching it holds that $\frac{f_{(T|M)}(T=0|M=1)}{f_{(T|M)}(T=1|M=1)} = \frac{k/(k+1)}{1/(k+1)} = k$, thus we can write (12) as

$$
w_i^c(\mathbf{x}) = \omega_i^t(\mathbf{x}) \times \frac{f_{(T|\mathbf{X},M)}(T=1|\mathbf{X}=\mathbf{x},M=1)}{1-f_{(T|\mathbf{X},M)}(T=1|\mathbf{X}=\mathbf{x},M=1)} \times k
$$

Also notice that for large matched sample it should hold that

$$
\frac{f_{(T|\mathbf{X},M)}(T=1|\mathbf{X}=\mathbf{x},M=1)}{1-f_{(T|\mathbf{X},M)}(T=1|\mathbf{X}=\mathbf{x},M)} = \frac{1}{k}
$$

Thus we can conclude that

$$
\omega_i^c(\mathbf{x}) = \omega_i^t(\mathbf{x})
$$

Therefore the weights that the units in the comparison group should be assigned for the outcome analysis are the same weight as the treated unit they have been matched to. Therefore the weights of the units in the comparison group are different from their original survey weights. Interestingly, when this weight transfer is implemented and then the PATT is estimated, we find that the estimation procedure assigns to each unit of the comparison group a final weight that is proportional to the final weight received by the treated unit to which they have been matched to. Such proportion is defined by the number of comparison units used to matched each treated unit. Notice that a similar result is obtained when considering simple random samples, see Stuart [2010].

3 Simulation Study

As we mentioned, the main goal of this article is to explore different methods used in the literature to estimate the PATT and compare their performance, under different non-response mechanisms. In the previous sections we argue that survey weights do not need to be incorporated in the estimation of the propensity score model and showed that survey weights of the units in the treated group should be transferred to the comparison units they have been matched to. In order to explore the empirical implications of these results we implement a simulation study in order to asses whether: (1) the performance of the propensity score matching estimator is affected by how (or if) the survey weights are incorporated in the estimation of the propensity score model, (2) the implementation of the weight transfer presented in Section 2.3 translates in a better performance of the propensity score matching estimator, and (3) our conclusions depend on the assumed non-response mechanism considered and on the difference between the SATT and the PATT

Our simulation set-up follows closely the one used by Austin et al. [2016], nevertheless we introduce some modifications that will be explicitly mentioned. As in Austin et al. [2016], we consider the case of a population of size $N = 1,000,000$. There are ten strata in the population, each stratum has a total of $100,000$ observations. Within each strata there are 20 clusters, each of is composed of $5,000$ units. There are **six covariates** X_l with $l = 1, ..., 6$ and the data generating mechanism for the baseline covariates is such that: (1) the probability density function is normal, (2) the covariates are independent (i.e., correlation between any pair of covariates is set equal to 0), (3) the standard deviation, across all the covariates, is equal to 1 and (4) the means vary across strata and cluster. More explicitly, for each strata (j), the mean of the covariates deviates in μ_{lj} from 0, where μ_{lj} are obtained assuming that $\mu_{lj} \sim N\left(0, \tau^{stratum}\right)$. Within each strata, the mean of each cluster (k) deviates from the strata specific mean by μ_{lk}, with $\mu_{lk} \sim N\left(0, \tau^{cluster}\right)$. Thus the distribution of the l^{th} variable, in the j^{th} stratum, among the units of the k^{th} cluster is $X_{l,ijk} \sim N\left(\mu_{lj} + \mu_{lk},\ 1\right)$. We set $\tau^{stratum} = 0.35$ and $\tau^{cluster} = 0.25,\ 0.15,\ 0.05$. Each value of $\tau^{cluster}$ defines a different scenario. Unless otherwise specified, values of the population coefficients are the ones used by Austin et al. [2016]

The **treatment assignment** (T_i) model is defined as a Bernoulli random variable $T_i \sim Be\left(p_i\right)$ with $logit\left(p_i\right) = \alpha_0 + \sum_{l=1}^{6} \alpha_l X_{l,i}$ with $\alpha_0 = log\left(\frac{0.3290}{0.9671}\right)$, $\alpha_1 = log(1.10)$, $\alpha_2 = log(1.25)$, $\alpha_3 = log(1.50)$, $\alpha_4 = log(1.75)$, $\alpha_5 = log(2.00)$ and $\alpha_6 = log(2.50)$.

The **potential outcomes** models are defined as $Y_i\left(0\right) = \beta_0 + \sum_{l=1}^{6} \beta_l X_{1,i} + \epsilon$ with $\epsilon \sim N(0,1)$ and $\beta_0 = 0$, $\beta_1 = 2.50$, $\beta_1 = -2.00$, $\beta_3 = 1.75$, $\beta_4 = -1.25$, and $\beta_6 = 1.10$. The potential outcome under treatment is defined by

$$Y_i\left(1\right) = Y_i\left(0\right) + \delta_0 + \delta_1 \sum_{l=1}^{3} \beta_l X_{l,i} + \sum_{j=1}^{10} \eta_j STR_{j,i}$$

with $\delta_0 = 1$ and $\delta_1 = 0.2$. The term $\sum_{j=1}^{10} \eta_j STR_{j,i}$ is the first departure form the simulation set-up design by Austin et al. [2016]. This additional term allows us to control how different the PATT and the SATT are. The variable $STR_{j,i}$ is a categorical variable that takes the

value 1 if the sample unit i belongs to the j^{th} stratum. For each of the three scenarios we consider six different values for the vector of parameters $(\eta_1, ..., \eta_{10})$ such that $\left(\frac{SATT}{PATT} - 1\right) \times 100$ takes roughly the values $-50\%, -40\%, -30\%, -20\%, -10\%$ and 0% . In addition to a continuous outcome Austin et al. [2016] also considered a dichotomous outcome; in our article, we restrict our attention to continuous outcomes.

In theirs simulation set-up,Austin et al. [2016] assumed that non-response rate was 0%. We extend the original simulation study and we consider 4 **non-response scenarios**. We define an indicator variable R_m with $m = 1, 2, 3, 4$ which takes the value 1 is the unit responded and 0 otherwise.

- No-missing data (**NM**): $R_{1i} = 1$ for all i.

- Missing at Random (**MAR**): the non-response rate depends on the six baseline covariates. Explicitly we assume that $R_{3i} \sim Be(p_{3i})$ with $logit\,(p_{3i}) = \gamma_0 + \sum_{l=1}^{6} \gamma_l X_{l,i}$ and $\gamma_0 = -log(0.030)$, $\gamma_1 = -log(1.10)$, $\gamma_2 = -log(1.25)$, $\gamma_3 = -log(1.50)$, $\gamma_4 = -log(1.75)$, $\gamma_5 = -log(2.00)$, $\gamma_6 = -log(2.50)$.

- Missing at Random with an additional covariate X_7 (**MARX**): the non-response rate depends on the baseline covariates but additionally, depends on a covariate X_7 that is not observed in the final sample, but affect the response rate. Formally,$R_{3i} \sim Be(p_{3i})$ with $logit\,(p_{3i}) = \gamma_0 + \sum_{l=1}^{7} \gamma_l X_{l,i}$ where $\gamma_7 = -log(2.50)$. This non-response mechanism, aims to model the situation in which the survey weights can are constructed using information that is only available to the survey team (i.e., X_7), but not available to the final user (e.g., number of contact attempts). The data generating mechanism for the covariate X_7 is the same as the one for the baseline covariates (i.e., $X_1, ..., X_6$)

- Missing at Random where the non-response depends on the baseline covariates and the treatment assignment (**MART**). Explicitly $R_{4i} \sim Be(p_{4i})$ with $logit\,(p_{3i}) = \gamma_0 + \sum_{l=1}^{6} \gamma_l X_{l,i} + \Delta T_i$ and $\Delta = -2$.

The final survey weights are defined as the number of individuals each person represents times the inverse of the probability of responding (we follow Little and Vartivarian [2003]). The average response rate across the MAR, MARX and MART models is close to 90%. We are aware that this response rate is high, nevertheless this allows us to compare the performance of the different PATT estimators without incurring in sample size adjustments. In general, samples sizes are increased by the inverse of the average response rate. By considering a relatively high respond rate, we don't need to implement such adjustments. We believe that increasing the non-response rate will only exacerbate the results that we observe in this article.

For each scenario and for each value of $\left(\frac{SATT}{PATT} - 1\right) \times 100$, we ran 1000 iterations, and compare the performance of different estimators. Throughout this article, performance is quantified by three metrics: (1) bias (in absolute value), (2) root mean square error (RMSE, which is defined as the sum of the squared bias and the variance associated with the estimator) and (3) empirical coverage of the 95% confidence interval.

3.1 Estimators of the PATT

The estimators of the PATT considered in our article are grouped based on: (1) how the survey weights are used in the estimation of the propensity score and (2) whether the weights transfer described in Section 2.3 is implemented.

Regarding the estimation of the propensity score, there are three alternatives to consider regarding the use of the survey weights: (1) not incorporate the weights in the estimation (**UPS**), (2) incorporate the weights in a weighted estimation (**WPS**), and (3) incorporate the survey weights as a covariate in the estimation of the propensity score model (**CPS**). Once the propensity score is estimated, a 1 to 1 matching will be implemented. We limit our attention to matching without replacement. After the matching procedure is executed, the survey weights of the treated can be transferred to the comparison units they have been matched to (**WT**) or each observation can retain their original survey weights (**OW**). Therefore, the estimator labeled as "**CPS|WT**" is the estimator of the PATT in which the survey weights are used as a covariate in the estimation of the propensity score model and the weight transfer described in Section 2.3 is implemented.

In addition to the 6 estimators previously described, we also consider a "**Naïve**" estimator of the PATT. The Naïve estimator uses propensity score matching but ignores the sampling design all together. That is, it does not incorporate the survey weights in the estimation of the propensity score nor uses them to weight the observed outcome, in other words the Naïve estimator is a valid estimator of the SATT.

Finally, following Ridgeway et al. [2015] we consider two outcome models: (1) an "unadjusted" model that has the treatment assignment (T) as the only regressor and (2) an "adjusted" model that in addition to the treatment assignment, includes the baselines covariates as regressors, but it does not include interaction terms. Is worth noticing that since simulated data using models with heterogeneous treatment effects, both outcome models considered are misspecified.

3.2 Results

Diagnostics

First we evaluate how balanced the distribution of the survey weights and the baseline covariates (i.e., $X_1, ..., X_6$) is between the treated and comparison groups as a result of implementing the matching procedures described in Section 3.1. Balance is defined in terms of the standardized mean difference (SMD). To facilitate comparisons in the ability to generate balanced samples across methods, we use a common standard deviation across all matching procedures. Notice that the SMD under the Naive estimation approach, provides a measure of balance achieved in the sample (not in the population) since it does not incorporate the survey weights. Since the other methods do incorporate survey weights in their estimation of causal effects, the calculation of the SMD associated with these estimators uses the survey weights, therefore we consider them measures of balance at the population level. Table 1, shows the SMD in the population, before any matching procedure was implemented.

Figures 1, 2 and 3 summarize our main findings for scenarios 1, 2 and 3, respectively. In these figures, the vertical axis displays the values of SMD computed in the matched sample. Each row of plots represent a different non-response scenarios described in Section 3.1. Each color represents the different procedure used in the estimation of the propensity score model: (1) purple is associated with the estimators that do not incorporate the survey weights in the estimation of the propensity score model, but the sample weights are used in the computation of the SMD after matching, (2) green bars represent the SMD after matching when the survey weights were incorporated in a weighted estimation of the propensity score model and (3) blue bars show the balance achieved after the matching procedure when the weights were used as a covariate in the estimation of the propensity score model. Darker shaded bars are associated with the implementation of the weight transfer described in Section2.3, lighter shades show the balance achieved (in terms of SMD) when the each sample unit kept its original sampling weight. We also display the SMD achieved by the Naive estimator in orange. The red line in Figure 1, shows the threshold value of 0.20 (see Rosenbaum and Rubin [1985]), if after matching the obtained SMD is greater than 0.20 we can conclude that the matching procedure was not effective in balancing the baseline covariates between the treated and comparison groups.

The patterns that we observe in Figure 1 are consistent across all scenarios (see Figures 2 and 3). First we observe that, in general, good balance is achieved by all matching procedures (there are some exceptions, the SMD for covariate X_6 is not always below 0.20 for some of the matching procedures when the non-response mechanism is MART). Secondly, when the non-response mechanism is either MAR, MARX or No Missing, the weight transfer described in Section 2.3 may translate into worse balance (this is particularly clear for covariate X_3 and in multiple covariates in Scenario 3). Nevertheless, notice that this situation is reversed when the non-response mechanism is MART. In fact, failure to implement the weight transfer described in Section 2.3 can generate improper balance in some of the covariates. Interestingly we observe that when the non-response mechanism is different from MART, balance in the covariates translates into balance of the survey weights.

Finally note that for most of the baseline covariates and across non-response mechanisms the Naive method achieves better balance that any other of the matching procedures implemented. Nevertheless, it is important to notice that the fact that the Naive method achieves good balance in the sample, does not imply the good balance is achieved by this procedure at the population level.

Treatment Effect Estimation Results

The estimators that only used the treatment indicator (T) as a covariate in the outcome model estimation are labeled as "$\mathbf{Y} \sim \mathbf{T}$", whereas the estimators that additionally adjust for the vector of covariates $\mathbf{X}$ are labeled as "$\mathbf{Y} \sim \mathbf{T} + \mathbf{X}$"

Figure 4 displays our findings for Scenario 1. Each column in the plot shows one of the metrics (bias in absolute value, empirical coverage of the 95% confidence interval and RMSE) chosen to asses the performance of the different matching estimators. Each row of plots represent a different non-response scenarios described in Section 3.1. We keep the same

color scheme used to asses the balance. The type of line is associated with how the survey weights we incorporated in the outcome estimation. Solid, darker lines are associated with the implementation of the weight transfer described in Section2.3, whereas dashed lighter lines show the performance achieved when each sample unit kept its original sampling weight. We also display the performance achieved by the Naive estimator in solid orange lines.

Differences in the performance of the estimators are more pronounced when we consider the "unadjusted" estimators. Notice that using the unadjusted outcome model to estimate the PATT is essentially the same procedure as the one described in Appendix A. As expected, as the percentage difference between the SATT and the PATT increases in absolute value, the naive estimator performance worsens. Notice that this results hold, even after adjusting for relevant covariates. As survey weights are incorporated in the analysis we find that keeping the original weights translates to reduction of bias (this is true for most of non-response mechanisms considered with the exception of MART). Adjusting for relevant covariates translates into better performance (across the three metrics considered). In general we observe that how the survey weights are incorporated in the estimation of the propensity score does not translate to a significant improvement in the performance of the estimators. When the non-response model is MART we observe that the weight transfer described in Section 2.3 reduces bias associated with the estimation of the PATT, this is true even after adjusting for relevant covariates (although is more explicit among the "unadjusted" estimators of the PATT). Furthermore, among the "unadjusted" estimators, we observe that the weight transfer is not only associated with better balance but also better coverage and better RMSE. Among the "unadjusted" estimators and when the non-response mechanisms is MART, we also observe that a weighted estimation of the propensity score models translates into gains of efficiency (reduction of the RMSE). Nevertheless, this gains do not seem to be significant when other relevant covariates are incorporated in the outcome model. We believe that the reason why the weight transfer described in Section 2.3 do not improve the performance of the estimators in other non-responses mechanisms (i.e., MAR, MARX and No-Missing) is due to the fact that if the matching procedure is successful, then balance in the covariates, will translate in balance of the survey weights and therefore the weight transfer is implicitly implemented. This hypothesis seems to be confirmed by Figures 1, 2 and 3, where we can observe that when that non-response mechanism is different from MART (i.e., the non-response is not a function of the treatment indicator), balance in the covariates translates into balance of the survey weights. Furthermore, notice that when the non-response is MART good balance of the baseline covariates does not translate into good balance of the survey weights, and therefore the weight transfer improves the performance of the estimators. Another key feature of the results depicted in Figure 4 (and also true in Figures 5 and 6), is that even when the percentage difference between the SATT and the PATT is as high as 50%, incorporating the survey weights translate into significant reduction of bias. It is also important to notice that the percentage difference between the SATT and the PATT gets close to 0, no significant differences in the performance of the naive and the rest of the estimators is observed (this is the default scenario of the simulation set-up implemented by Austin et al. [2016]).

Analyzing the balance achieved and the performance of each method, we can conclude that when estimating the PATT using complex survey data, the **population** balance (measured as SMD that incorporates the survey weights in its computation) helps to predict the performance of the different estimators (therefore the balance achieved by the naive method can be misleading). This can be seen in Figure 6 where the ability of estimator labeled "**CPS|WT**", to achieve a better balance across all the covariates considered (when the non-response mechanism is MART) translates into a better performance in terms of bias. In general we conclude that the survey weights need to be included (at least) in the computation of balance measures and the outcome analysis when estimating the PATT using complex survey data.

4 Application

In this section we use The Early Childhood Longitudinal Study, Kindergarten class 1998-1999 (ECLS-K) (see [Tourangeau et al., 2009]) to estimate the effect of special education services in math skills. Furthermore, we try to replicate the results by Keller and Tipton [2016]. Keller and Tipton [2016] provide an excellent guide on how to implement different R packages to estimate causal effects by implementing different matching procedures. The authors illustrate how different packages work by replicating the work of Morgan et al. [2008]. We follow closely the work by Keller and Tipton [2016] since they provide a comprehensive list of the variables used in their analysis. It is worth noticing that Morgan et al. [2008] does not explicitly mention how the survey weights are incorporated in the estimation of the propensity score matching estimators and neither does Keller and Tipton [2016] (although Keller and Tipton [2016] explicitly state that the purpose of their article is to illustrate how different software can be use to implement propensity score matching and their results should not be interpreted in a causal context).

The ECLS-K is a longitudinal study that examines child early school experiences beginning with kindergarten until eight grade. The ECLS-K is a nationally representative sample that collects information: (1) at the child level (including academic information, cognitive, emotional, physical and social development, etc), (2) at the household level (socio-economical status, household composition, parents education, etc) and (3) at the school level (educational practices, school socio-demographics, curriculum, teachers' qualifications, etc). Throughout this application we use as survey weights the variable labeled as `"C1_6FC0"` in the ECLS-K data set. By design, the ECLS-K sample did not follow students who transfer from their original school. Thus the final sample has been re-weighted such that the reminder of the observations are representative of the kindergarten population (see Reardon et al. [2009]). The data was accessed through ResearchConnections.org (see U.S. Department of Education. Institute of Education Sciences. National Center for Education Statistics. [2011])

Since our goal is to illustrate the methods we do not assess the plausibility of the key assumptions that would be needed to interpret the results as causal in this particular application, and thus the results should not be treated as definitive causal effects regarding the effect of special of education services on learning skills.Morgan et al. [2008], evaluate

the effectiveness of special education services on: (1) math and reading skills, (2) learning related behaviors and (3) behavioral problems. For the sake of simplicity we follow Keller and Tipton [2016] and focus only on the average effect of elementary school special education services (i.e., the exposure) on math achievement in fifth grade (i.e., the observed outcome). We use the of variables listed in Keller and Tipton [2016] to estimate the propensity score model. We consider 39 covariates in the propensity score model (which include demographic, socio-economical, academic, household and school level variables). A codebook of the variables used in this application is available as supplementary material. We fit an unadjusted outcome model, that is we regress the observed outcome (i.e., math skills) on the treatment or exposure (i.e., whether or not the child in question, receives special education services) without incorporating other covariates. Additionally, we also fit an outcome model adjusting for the same set of covariates considered in the propensity score model. The first column in Table 2 shows the SMD computed without using the survey weights (either in the estimation of the propensity score model or in the computation of the SMD). Overall we observe that most of the matching procedures were effective in increasing the balance for this set of covariates. Nevertheless, some of the methods were not able to improve the balance enough to generate SMDs smaller than 0.20 in some of the covariates (see the highlighted cells in Table 2). Notice that **WPS|WT** is the only method that has achieved SMDs smaller that 0.20 in 38 of the 39 covariates considered. The last row in Table 2 show the SMD of the survey weights after the matching procedure. Notice that the good balance in the covariates does translate into good balance in the survey weights, this result seems to indicate that the survey weights may not be affected by the treatment indicator, therefore weight transfer derived in Section 2.3 may not improve the performance of the matching estimator. Furthermore we observe more volatility among those estimators that implement such transfer.

The next table (Table 3), shows the estimated PATT. The first column displays the estimation result of implementing an unadjusted regression model and the second column shows the associated 95% confidence interval. The third column, shows the results of estimating the PATT adjusting for the set of covariates considered in Table 2, and the last column shows the associated 95% confidence interval. As expected most of the estimators produce similar estimators with the exception of the Naive estimator.

5 Discussion

In this article we explore how different ways of utilizing survey weights can affect the performance of propensity score matching estimators of the PATT in the context of complex survey data, when different non-response mechanism are considered. We hope that this article provides some insight about what are the best practices when it comes to estimate the PATT with complex survey data. There has been only limited work to extend propensity score matching techniques in the context complex survey data, we hope that this article will help to reduce such gap.

To our knowledge, this is the first article that explores the how different non-response mechanisms can affect the performance of different propensity score matching estimators.

This a key extension, since non-response tends to be the rule rather the exception in the context of complex survey data.

We have also evaluated how differences in that SATT and the PATT affect the performance of different propensity score matching estimators. We were able to address this issue by extending the outcome model proposed by Austin et al. [2016] in their simulation study, by adding strata specific effects. The rationale to incorporate this additional terms, is based on the fact that when we first replicated the simulation study designed by Austin et al. [2016] we found that the naïve estimator of the PATT (i.e., an estimator that does not incorporate the survey weights in the estimation of the propensity score nor as weights of the observed outcome) performed as well as any of the other estimators considered by Austin et al. [2016]. The reason why the naïve estimator had such a good performance was due to the fact that the PATT and the SATT where practically identical, in fact the percentage difference between that PATT and that SATT (defined as the average of $\left(\frac{SATT}{PATT} - 1\right) \times 100$) was less that 1%. Therefore it was not possible to assess whether the observed performance of the different estimators considered by Austin et al. [2016] was due to the fact that the SATT and PATT where so similar or to the fact that it did not matter how the survey weights were used in the estimation of the propensity score. We hope that hour extension will help to elucidate how survey weights should be incorporated even in situation where the SATT and the PATT are drastically different.

Based on the data motivated application using the ECLS-K data set and results of the simulation study we conclude that:

- **How the survey weights are incorporated in the estimation of the propensity score, does not affect the performance of the matching estimators.** In our simulation study we used the survey weights in three different ways: (1) the weights were incorporated as a covariate in the estimation of the propensity score model,(2) they were incorporated as weights in a weighted regression analysis, and (3) they were not used at all in the estimation of the propensity score. We found that how the survey weights were incorporated in the estimation of the propensity score did not help to predict the performance of the estimators. This result holds true across for all non-response mechanisms, although we found evidence that a weighted estimation of the propensity score model can increase the efficiency of the PATT estimator when an unadjusted outcome model is estimated and the missing data pattern is MART. Some of these results are consistent with those from Austin et al. [2016]

- **Adjusting for relevant covariates in the outcome model improves the performance of the estimators.** We found that adjusting for relevant covariates in the outcome analysis translates into better performance of the estimators in general. This results are consistent with the ones found by Ridgeway et al. [2015]. In the literature there are multiple resources that explore the consequences of misspecified models (either the propensity score or the outcome model) when estimating causal effects. See Drake [1993], McCaffrey et al. [2004], Frölich [2007], Robins et al. [2007], Lee et al. [2011] and Imai and Ratkovic [2014], among others.

- **survey weights should be incorporated in the outcome analysis.** When survey weights are incorporated in the outcome analysis we find that performance of estimators that include relevant covariates and include the survey weights in the outcome analysis have the best performance. Our results indicate that not including survey weights in the estimation procedure may lead to substantial bias, specially when the SATT and the PATT are different.

- **Weight transfer improves the performance of the matching estimators under certain non-response mechanisms.** We find that the weight transfer described in Section 2.3 helps to improve the performance of the estimators specially when the missing data mechanism is MART (i.e., depends on the treatment status) greater reduction in bias are observed when the PATT is estimated using a unadjusted outcome model.

- **Balance is crucial to correctly estimate treatment effects using propensity score matching.** We found that the key element to obtain estimators of the PATT that have a good performance, is to achieve good **population** balance in the observed covariates. That is, survey weights need to be incorporated in the computation of the measure of balance. Balance was the best predictor of the performance of the PATT estimator. In fact, from our simulation study we observe that the average correlation (across the covariates) between bias and SMD achieved by the estimators that include survey weights in their estimation (that is, excluding the naive estimator) is 0.77. And the correlations (also excluding the naive estimator) between Coverage and RMSE are -0.66 and 0.62 respectively. We also computed the correlation between the different metrics of performance (i.e., Bias, Coverage and RMSE) and the SMD achieved by Naive estimator. The correlation between bias and SMD for the naive estimator is 0.00, between coverage and SMD is -0.15 and between RMSE and SMD is -0.1. Therefore we can conclude that good sample balance does not, necessarily translate into good performance of the propensity score matching estimator.

- **The balance achieved in the survey weights after the matching procedure, could potentially help identify the nature of the non-response mechanism.** From our simulation study and under the assumption of no unmeasured confounders, we noticed that when the non-response is different from MART good balance in the baseline covariates translates into good balance of the survey weights, thus the weight transfer described in Section 2.3 is implicitly implemented (hence we observe similar performances of the different propensity score matching estimators). Nevertheless, when the non-response mechanism is MART, good balance in the confounders does not imply balance of the survey weights, and thus the weight transfer improves the performance of the estimators considered.

Given our results we recommend that, before estimating the PATT, **population** balance of the relevant covariates should be evaluated. In other words, survey weights should be incorporated in the computation of the selected balance measures. We suggest that if proper

balance is not achieved, the propensity score model should be refined until satisfactory balance among the observed covariates is attained. Nevertheless, notice that this refinement procedure will be hard to incorporate in the final inference.

It is important to note that the confidence intervals presented in this article were constructed using the 'survey' package in R (see Lumley [2016]). There has been little work to evaluate the asymptotic properties of matching estimators, with the exception of significant contributions made by Abadie and Imbens [2006], Abadie and Imbens [2008] and Abadie and Imbens [2016]. In these articles the authors not only evaluate consistency of the matching estimator, but evaluate the performance of common practices (i.e., estimating standard errors of matching estimators by implementing bootstrap) and derive the asymptotic distribution of matching estimators when the estimated propensity score is used to create the matched sample. Abadie and Imbens [2016] derive the asymptotic distribution of the matching estimator in the context of simple random sampling. Future work will focus on generalizing those results to the context of complex survey data.

In our article we have restricted our attention to matching estimators of the PATT, where the matching procedure was implemented without replacement. In the context of simple random samples, when matching is estimated with replacement weights should created to guarantee that the matched treated and comparison groups are weighted up to be similar (see Ho et al. [2011]), future work will consist in extending the computation of such weights in the context of complex survey data.

Lastly, in our simulation study we have implicitly assume that the propensity score model was correctly specified. Future work will evaluate how the performance of propensity score matching estimators is affected by the degree of misspecification of both, the propensity score model and the outcome model.

Acknowledgments

The authors wish to thank Francis M. Abreu, whose comments greatly improved this manuscript. The research reported here was supported in part by the Institute of Education Sciences, U.S. Department of Education, through grant R305D150001 (PI: Stuart).

Software

In this article the simulation study, plots and the application were implemented using the software R and the platform RStudio (see RStudio Team [2015]). The following packages were used: 'data.table' (see Dowle et al. [2015]), 'ggplot2' (see Wickham [2009]), 'MatchIt' (see Ho et al. [2011]), 'survey' (Lumley [2016]), sampling (see Tillé and Matei [2015]) and 'xtable' (Dahl [2016]). The code used to implement the simulation study, the application and the plots is available as supplementary material

References

A. Abadie and G. W. Imbens. Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74(1):235–267, 2006.

A. Abadie and G. W. Imbens. On the failure of the bootstrap for matching estimators. *Econometrica*, 76(6):1537–1557, 2008.

A. Abadie and G. W. Imbens. Matching on the estimated propensity score. *Econometrica*, 84(2):781–807, 2016.

P. C. Austin, N. Jembere, and M. Chiu. Propensity score matching and complex surveys. *Statistical Methods in Medical Research*, page 0962280216658920, 2016.

T. L. Brunell and J. DiNardo. A propensity score reweighting approach to estimating the partisan effects of full turnout in american presidential elections. *Political Analysis*, 12(1): 28–45, 2004.

W. G. Cochran. Sampling techniques. 1977. *New York: John Wiley and Sons*, 1977.

D. B. Dahl. *xtable: Export Tables to LaTeX or HTML*, 2016. URL `https://CRAN.R-project.org/package=xtable`. R package version 1.8-2.

M. Dowle, A. Srinivasan, T. Short, S. L. with contributions from R Saporta, and E. Antonyan. *data.table: Extension of Data.frame*, 2015. URL `https://CRAN.R-project.org/package=data.table`. R package version 1.9.6.

C. Drake. Effects of misspecification of the propensity score on estimators of treatment effect. *Biometrics*, pages 1231–1236, 1993.

M. Frölich. Finite-sample properties of propensity-score matching and weighting estimators. *Review of Economics and Statistics*, 86(1):77–90, 2004.

M. Frölich. Propensity score matching without conditional independence assumption—with an application to the gender wage gap in the united kingdom. *The Econometrics Journal*, 10(2):359–407, 2007.

A. Gelman. Struggles with survey weighting and regression modeling. *Statistical Science*, pages 153–164, 2007.

E. Grau, F. Potter, S. Williams, and N. Diaz-Tena. Nonresponse adjustment using logistic regression: To weight or not to weight. *American Statistical Association, Survey Research Methods Section. Alexandria*, pages 3073–3080, 2006.

R. M. Groves, F. J. Fowler Jr, M. P. Couper, J. M. Lepkowski, E. Singer, and R. Tourangeau. *Survey methodology*, volume 561. John Wiley & Sons, 2009.

D. L. Hahs-Vaughn and A. J. Onwuegbuzie. Estimating and using propensity score analysis with complex samples. *The Journal of Experimental Education*, 75(1):31–65, 2006.

K. Hallberg, P. M. Steiner, and T. D. Cook. The role of pretest and proxy-pretest measures of the outcome for removing selection bias in observational studies. *Society for Research on Educational Effectiveness*, 2011.

M. H. Hansen, W. G. Madow, and B. J. Tepping. An evaluation of model-dependent and probability-sampling inferences in sample surveys. *Journal of the American Statistical Association*, 78(384):776–793, 1983.

J. J. Heckman and P. E. Todd. A note on adapting propensity score matching and selection models to choice based samples. *The econometrics journal*, 12(s1):S230–S234, 2009.

M. A. Hernán and J. M. Robins. *Causal Inference*. Boca Raton: Chapman & Hall/CRC. Forthcoming, 2017.

D. E. Ho, K. Imai, G. King, and E. A. Stuart. MatchIt: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software*, 42(8):1–28, 2011. URL `http://www.jstatsoft.org/v42/i08/`.

G. Hong and S. W. Raudenbush. Effects of kindergarten retention policy on children's cognitive growth in reading and mathematics. *Educational Evaluation and Policy Analysis*, 27(3):205–224, 2005.

K. Imai and M. Ratkovic. Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):243–263, 2014.

B. Keller and E. Tipton. Propensity score analysis in ra software review. *Journal of Educational and Behavioral Statistics*, page 1076998616631744, 2016.

S. Korenman, K. S. Abner, R. Kaestner, and R. A. Gordon. The child and adult care food program and the nutrition of preschoolers. *Early childhood research quarterly*, 28(2): 325–336, 2013.

E. L. Korn and B. I. Graubard. Analysis of large health surveys: accounting for the sampling design. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, pages 263–295, 1995a.

E. L. Korn and B. I. Graubard. Examples of differing weighted and unweighted estimates from a sample survey. *The American Statistician*, 49(3):291–295, 1995b.

B. K. Lee, J. Lessler, and E. A. Stuart. Weight trimming and propensity score weighting. *PloS one*, 6(3):e18174, 2011.

R. Little. The bayesian approach to sample survey inference. *Analysis of Survey Data*, pages 49–57, 2003.

R. J. Little and S. Vartivarian. On weighting the rates in non-response weights. *Statistics in medicine*, 22(9):1589–1599, 2003.

T. Lumley. survey: analysis of complex survey samples, 2016. R package version 3.31.

D. F. McCaffrey, G. Ridgeway, and A. R. Morral. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological methods*, 9 (4):403, 2004.

P. L. Morgan, M. L. Frisco, G. Farkas, and J. Hibel. A propensity score matching analysis of the effects of special education services. *The Journal of special education*, 2008.

U. S. D. of Education. Institute of Education Sciences. National Center for Education Statistics. Early childhood longitudinal study [united states]: Kindergarten class of 1998-1999, kindergarten-eighth grade full sample. icpsr28023-v1. ann arbor, mi: Inter-university consortium for political and social research [distributor]. `http://doi.org/10.3886/ICPSR28023.v1http://doi.org/10.3886/ICPSR28023.v1`, 2011.

F. Potter, E. Grau, S. Williams, N. Diaz-Tena, and B. L. Carlson. An application of propensity modeling: Comparing unweighted and weighted logistic regression models for nonresponse adjustments. In *Proceedings of the Survey Research Methods Section. American Statistical Association*, 2006.

S. F. Reardon, J. E. Cheadle, and J. P. Robinson. The effect of catholic schooling on math and reading development in kindergarten through fifth grade. *Journal of Research on Educational Effectiveness*, 2(1):45–87, 2009.

G. Ridgeway, S. A. Kovalchik, B. A. Griffin, and M. U. Kabeto. Propensity score analysis with survey weighted data. *Journal of Causal Inference*, 3(2):237–249, 2015.

J. Robins, M. Sued, Q. Lei-Gomez, and A. Rotnitzky. Comment: Performance of double-robust estimators when" inverse probability" weights are highly variable. *Statistical Science*, 22(4):544–559, 2007.

P. R. Rosenbaum and D. B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

P. R. Rosenbaum and D. B. Rubin. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39 (1):33–38, 1985.

RStudio Team. *RStudio: Integrated Development Environment for R*. RStudio, Inc., Boston, MA, 2015. URL `http://www.rstudio.com/`.

D. B. Rubin. Matching to remove bias in observational studies. *Biometrics*, pages 159–183, 1973.

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

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

M. Schonlau, K. Zapert, L. P. Simon, K. H. Sanstad, S. M. Marcus, J. Adams, M. Spranca, H. Kan, R. Turner, and S. H. Berry. A comparison between responses from a propensity-weighted web survey and an identical rdd survey. *Social Science Computer Review*, 22(1): 128–138, 2004.

E. A. Stuart. Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 25(1):1, 2010.

Y. Tillé and A. Matei. *sampling: Survey Sampling*, 2015. URL `https://CRAN.R-project.org/package=sampling`. R package version 2.7.

K. Tourangeau, C. Nord, T. Lê, A. G. Sorongon, and M. Najarian. Early childhood longitudinal study, kindergarten class of 1998-99 (ecls-k): Combined user's manual for the ecls-k eighth-grade and k-8 full sample data files and electronic codebooks. nces 2009-004. *National Center for Education Statistics*, 2009.

H. Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2009. ISBN 978-0-387-98140-6. URL `http://ggplot2.org`.

E. Zanutto, B. Lu, and R. Hornik. Using propensity score subclassification for multiple treatment doses to evaluate a national antidrug media campaign. *Journal of Educational and Behavioral Statistics*, 30(1):59–73, 2005.

E. L. Zanutto. A comparison of propensity score and linear regression analysis of complex survey data. *Journal of data Science*, 4(1):67–91, 2006.

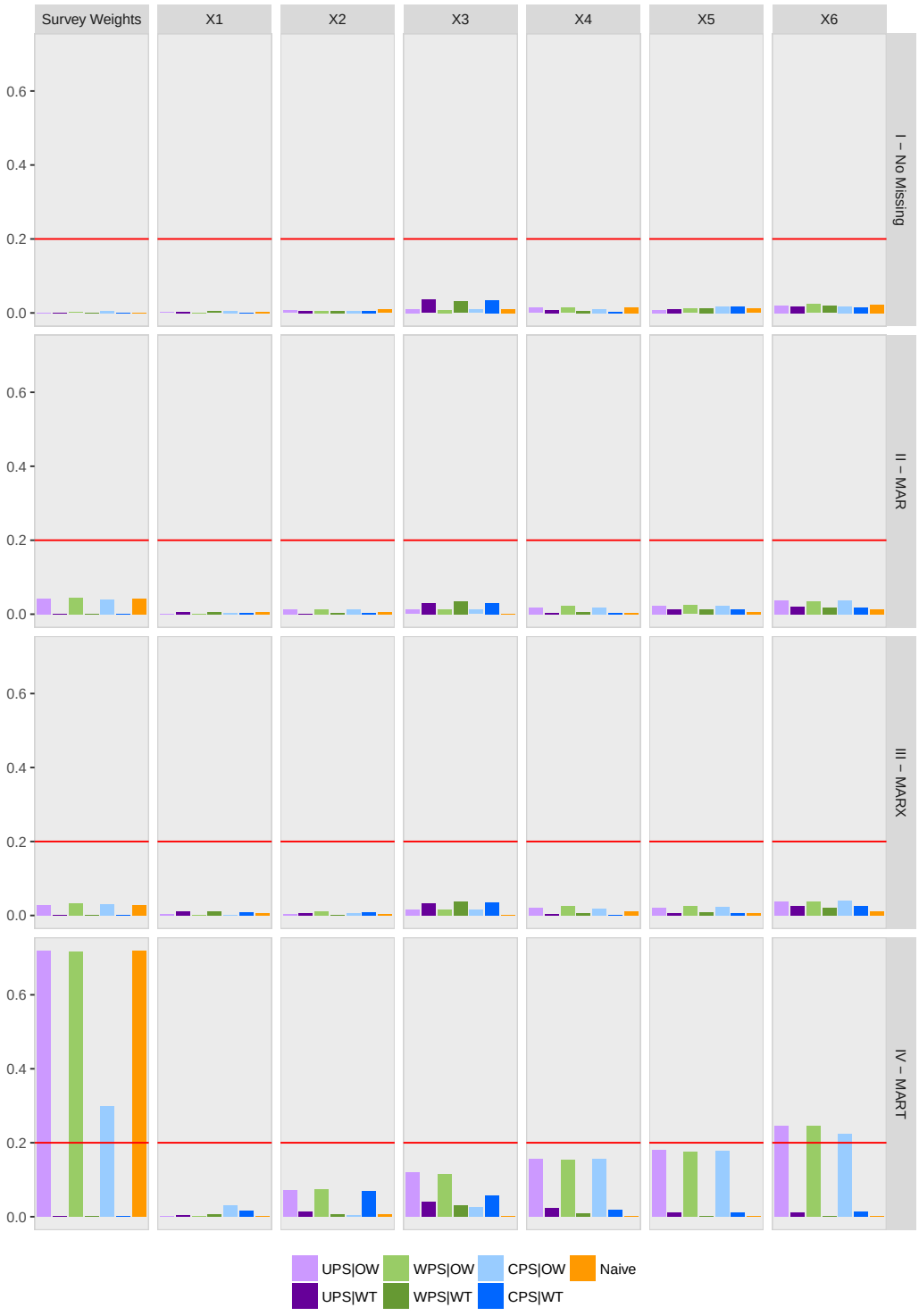

Figure 1: **Diagnostics.** SMD computed in the matched samples in Scenario 1. The Naive method represents balance in the sample (since it does not involve the survey weights in its computation). The other methods incorporate survey weights in the computation of the SMD, thus are considered measures of balance in the population.

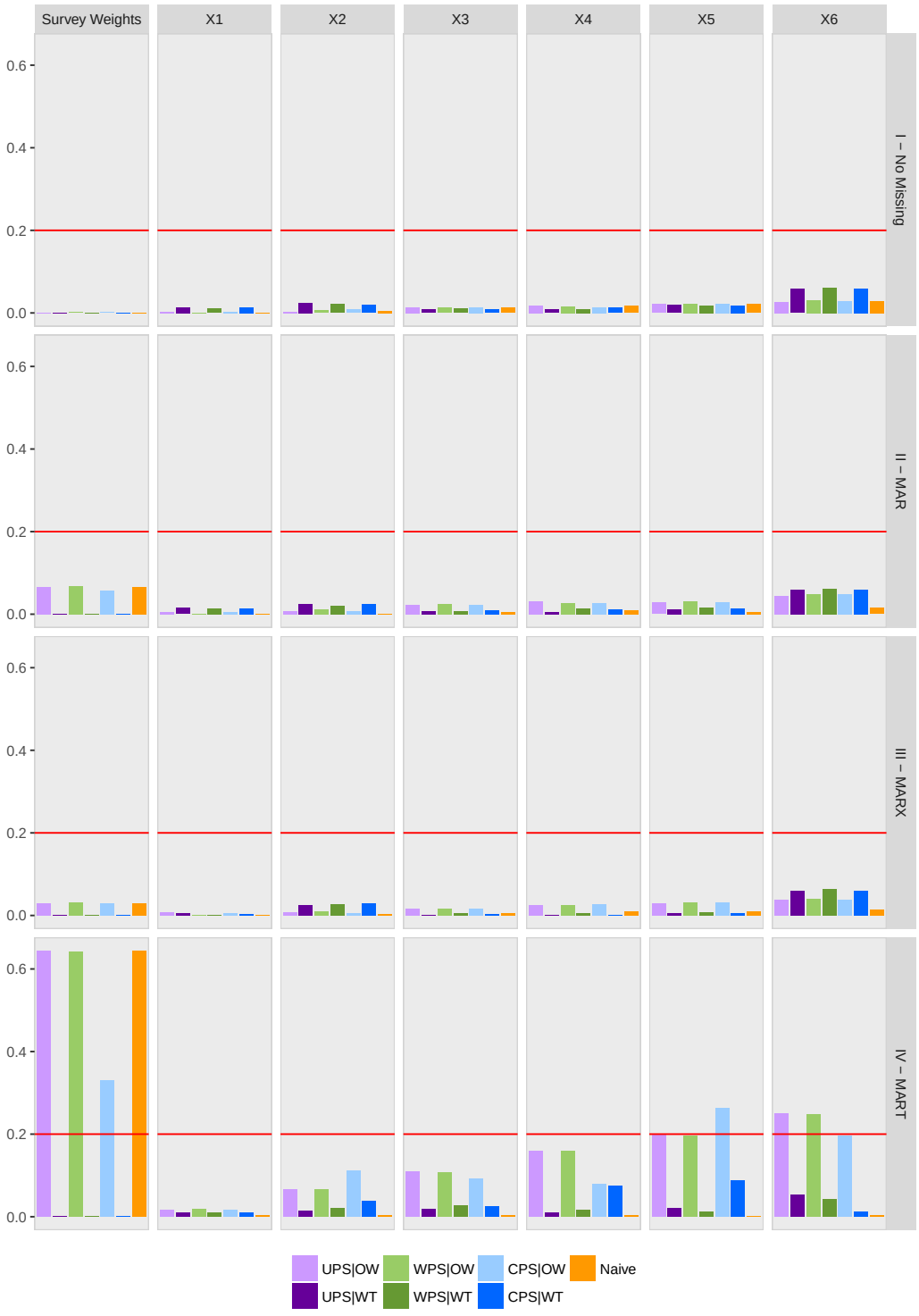

Figure 2: **Diagnostics.** SMD computed in the matched samples in Scenario 2. The Naive method represents balance in the sample (since it does not involve the survey weights in its computation). The other methods incorporate survey weights in the computation of the SMD, thus are considered measures of balance in the population.

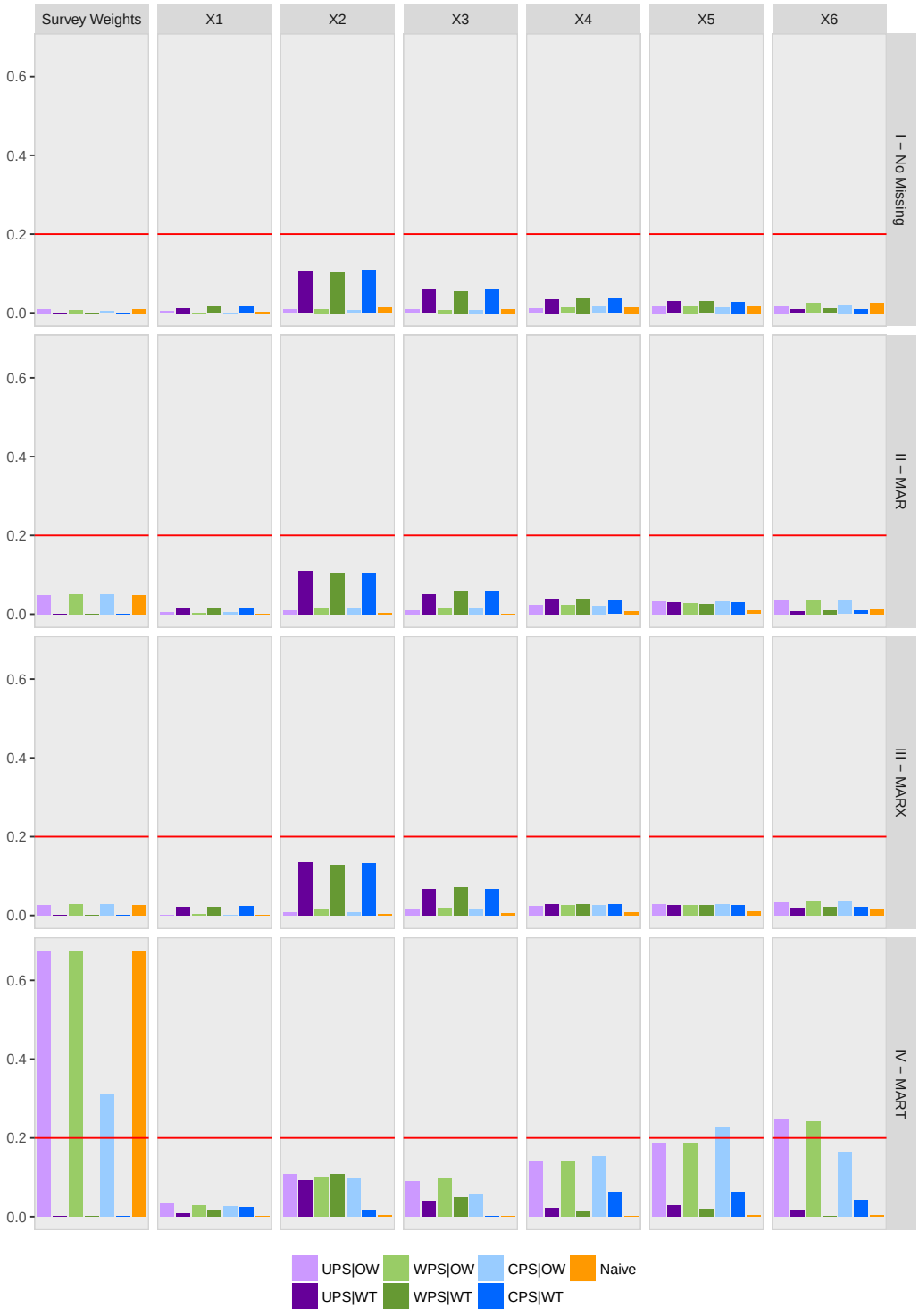

Figure 3: **Diagnostics.** SMD computed in the matched samples in Scenario 3. The Naive method represents balance in the sample (since it does not involve the survey weights in its computation). The other methods incorporate survey weights in the computation of the SMD, thus are considered measures of balance in the population.

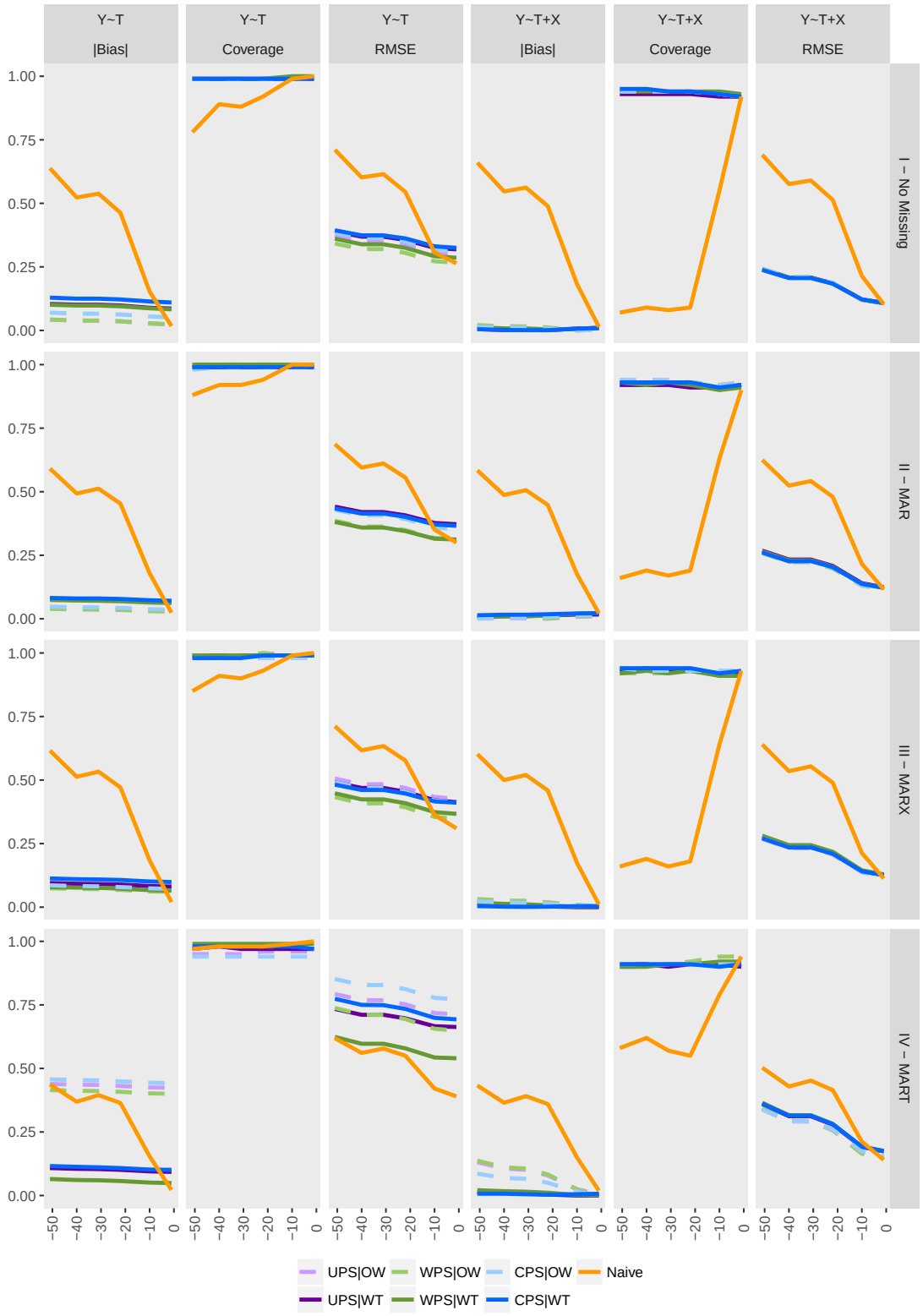

Figure 4: **Scenario 1** Bias in absolute value, coverage and root mean squared error (RMSE) as functions of the % difference between the SATT and PATT (simulation study).

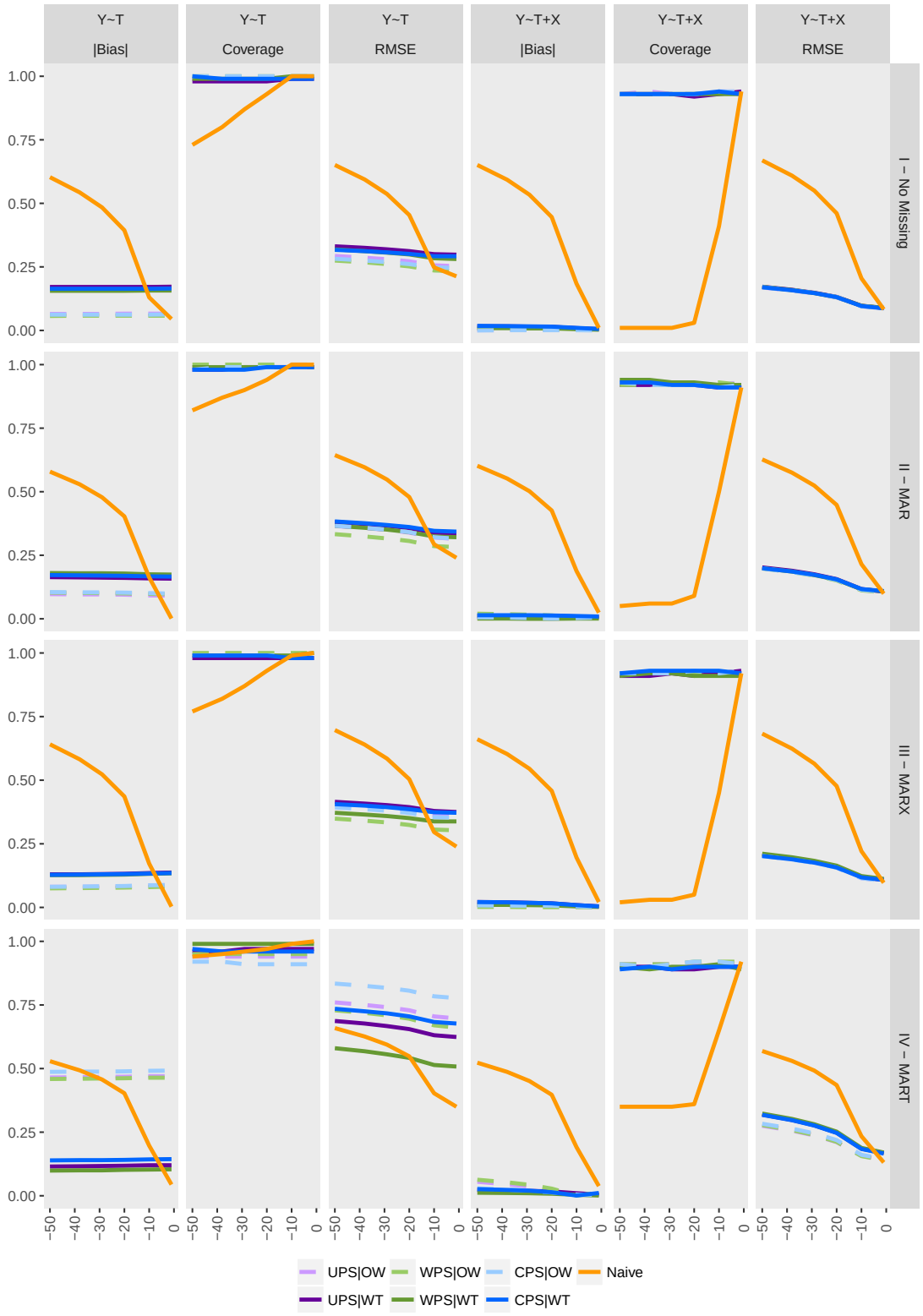

Figure 5: **Scenario 2** Bias in absolute value, coverage and root mean squared error (RMSE) as functions of the % difference between the SATT and PATT (simulation study).

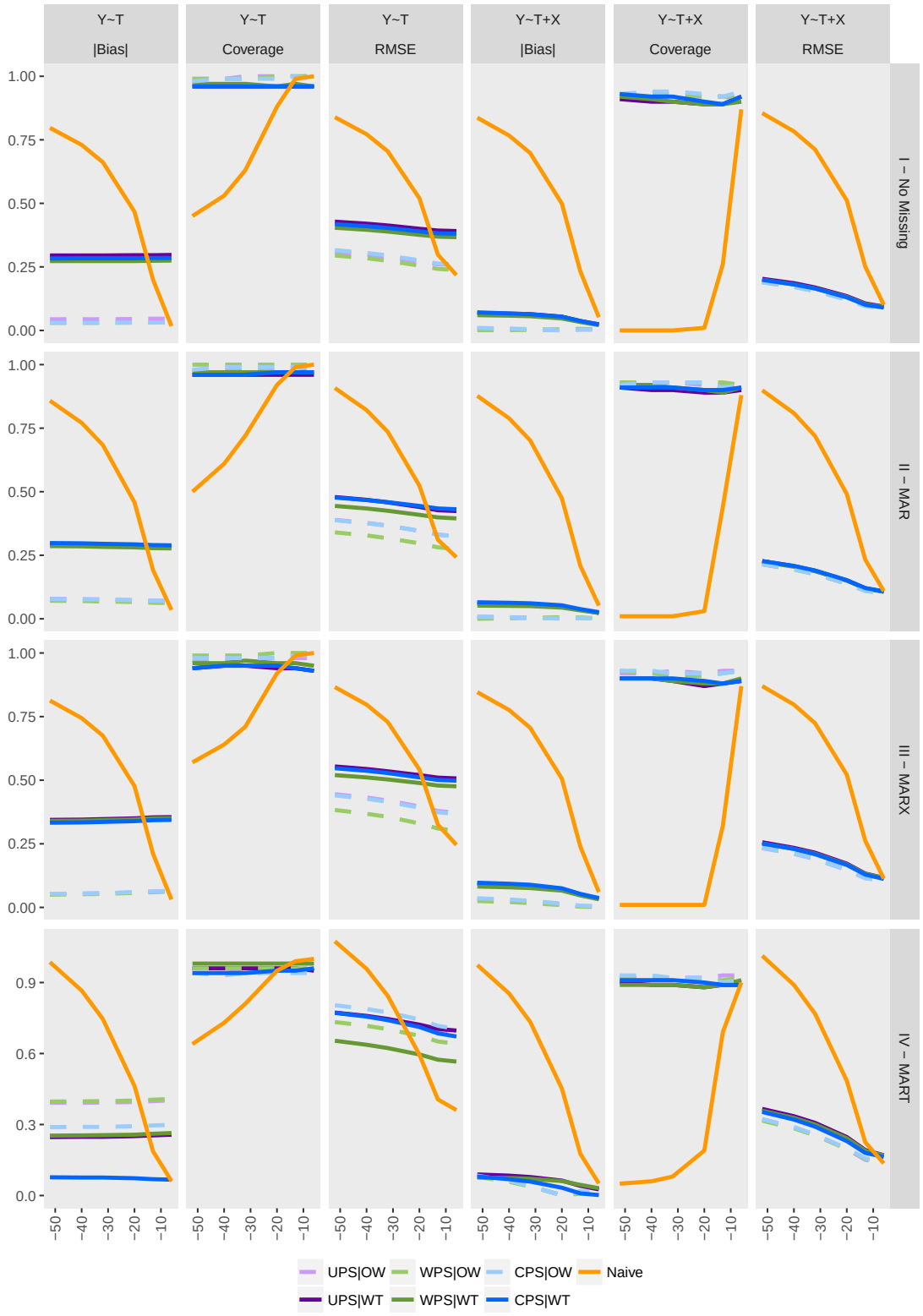

Figure 6: **Scenario 3** Bias in absolute value, coverage and root mean squared error (RMSE) as functions of the % difference between the SATT and PATT (simulation study).

Table 1: *Standardized Mean Differences (Population level)*

Scenario	X1	X2	X3	X4	X5	X6
1	0.11	0.28	0.43	0.49	0.69	0.81
2	0.03	0.16	0.34	0.59	0.57	0.91
3	0.09	0.22	0.33	0.48	0.60	0.81

Table 2: SMD achieved by the different estimation procedures.

Variable	Naive	UPS|OW	UPS|WT	CPS|OW	CPS|WT	WPS|OW	WPS|WT
FEMALE	0.08	0.08	0.06	0.08	0.06	0.03	0.01
WHITE	0.05	0.05	0.05	0.05	0.05	0.10	0.14
WKSESL	0.04	0.04	0.06	0.04	0.06	0.05	0.09
C1R4RSCL	0.04	0.04	0.03	0.04	0.03	0.06	0.04
C1R4MSCL	0.13	0.13	0.24	0.13	0.24	0.00	0.04
S2KPUPRI	0.25	0.25	0.15	0.25	0.15	0.01	0.02
P1ELHS	0.02	0.02	0.04	0.02	0.04	0.02	0.10
P1EHS	0.06	0.06	0.09	0.06	0.09	0.05	0.03
P1ESC	0.10	0.10	0.08	0.10	0.08	0.02	0.00
P1EC	0.15	0.15	0.09	0.15	0.09	0.13	0.04
P1EMS	0.00	0.00	0.07	0.00	0.07	0.08	0.04
P1EPHD	0.04	0.04	0.00	0.04	0.00	0.10	0.05
P1FIRKDG	0.16	0.16	0.14	0.16	0.14	0.20	0.17
P1AGEENT	0.12	0.12	0.07	0.12	0.07	0.06	0.06
T1LEARN	0.01	0.01	0.05	0.01	0.05	0.05	0.10
P1HSEVER	0.03	0.03	0.03	0.03	0.03	0.03	0.16
FKCHGSCH	0.00	0.00	0.09	0.00	0.09	0.05	0.12
S2KMINOR	0.10	0.10	0.09	0.10	0.09	0.21	0.12
P1FSTAMP	0.02	0.02	0.13	0.02	0.13	0.05	0.14
SGLPAR	0.05	0.05	0.07	0.05	0.07	0.04	0.17

Continued on next page

Table 2 – continued from previous page

Variable	Naive	UPS|OW	UPS|WT	CPS|OW	CPS|WT	WPS|OW	WPS|WT
TWOPAR	0.05	0.05	0.07	0.05	0.07	0.04	0.17
P1NUMSIB	0.06	0.06	0.01	0.06	0.01	0.07	0.12
P1HMAFB	0.04	0.04	0.17	0.04	0.17	0.03	0.19
WKCAREPK	0.03	0.03	0.14	0.03	0.14	0.06	0.06
P1EARLY	0.07	0.07	0.09	0.07	0.09	0.05	0.09
P1WEIGHO	0.06	0.06	0.11	0.06	0.11	0.05	0.09
C1FMOTOR	0.14	0.14	0.29	0.14	0.29	0.13	0.11
C1GMOTOR	0.15	0.15	0.20	0.15	0.20	0.06	0.07
P1HSCALE	0.12	0.12	0.08	0.12	0.08	0.04	0.05
P1SADLON	0.04	0.04	0.22	0.04	0.22	0.02	0.01
P1IMPULS	0.09	0.09	0.17	0.09	0.17	0.02	0.06
P1ATTENI	0.14	0.14	0.23	0.14	0.23	0.10	0.04
P1SOLVE	0.26	0.26	0.38	0.26	0.38	0.20	0.14
P1PRONOU	0.03	0.03	0.10	0.03	0.10	0.28	0.26
P1DISABL	0.13	0.13	0.08	0.13	0.08	0.12	0.04
AVG4RSCL	0.03	0.03	0.04	0.03	0.04	0.15	0.03
AVG4MSCL	0.01	0.01	0.04	0.01	0.04	0.19	0.02
AVGWKSES	0.03	0.03	0.06	0.03	0.06	0.14	0.03
C1_6FC0	0.11	0.11	0.00	0.11	0.00	0.08	0.00

Table 3: *PATT estimation. Unadjusted vs. Adjusted*

	Unadjusted	95% CI	Adjusted	95% CI
Naive	-2.62	(-4.44; -0.81)	-3.30	(-5.98; -0.61)
UPS|OW	-5.25	(-8.55; -1.94)	-7.86	(-13.42; -2.30)
UPS|WT	-4.33	(-7.24; -1.42)	-9.92	(-14.98; -4.86)
CPS|OW	-5.79	(-8.98; -2.61)	-6.63	(-12.18; -1.08)
CPS|WT	-5.31	(-8.39; -2.24)	-7.59	(-12.89; -2.29)
WPS|OW	-4.62	(-8.05; -1.19)	-6.39	(-11.90; -0.88)
WPS|WT	-2.80	(-6.13; 0.53)	-5.97	(-11.34; -0.61)

Appendix A Estimating the PATT

Here we incorporate the weight transfer derived in the previous Section 2.3 to estimate the PATT. The PATT will be estimated as the difference of the weighted mean of the observed outcomes of the treated and their matched comparison units. This estimator of the PATT makes the use of the weights explicit, nevertheless it is important to recall that a outcome model can be defined and the weights can be incorporated in its estimation. Under the assumption that a $k:1$ matching procedure was implemented it holds that for every treated unit j with $j = 1, 2, \ldots, n_T = \sum_{i=1}^{N} SR_i \times T_i$, we have $h(j) = 1, 2, \ldots k$ comparison units. Thus the $PATT$ can be computed by

$$\begin{aligned}
\widehat{PATT} &= \frac{\sum_{j=1}^{n_T} y_j \times \omega_j^t(\mathbf{x})}{\sum_{j=1}^{n_T} \omega_j^t(\mathbf{x})} - \frac{\sum_{j=1}^{n_T} \sum_{h(j)=1}^{k} y_{h(j)} \times \omega_j^t(\mathbf{x})}{\sum_{j=1}^{n_T} \sum_{h(j)=1}^{k} \omega_j^t(\mathbf{x})} \\
&= \frac{\sum_{j=1}^{n_T} y_j \times \omega_j^t(\mathbf{x})}{\sum_{j=1}^{n_T} \omega_j^t(\mathbf{x})} - \frac{\sum_{j=1}^{n_T} \sum_{h(j)=1}^{k} y_{h(j)} \times \omega_j^t(\mathbf{x})}{\sum_{h(1)=1}^{k} \omega_1^t(\mathbf{x}) + \sum_{h(2)=1}^{k} \omega_2^t(\mathbf{x}) + \ldots + \sum_{h(n_T)=1}^{k} \omega_{n_T}^t(\mathbf{x})} \\
&= \frac{\sum_{j=1}^{n_T} y_j \times \omega_j^t(\mathbf{x})}{\sum_{j=1}^{n_T} \omega_j^t(\mathbf{x})} - \frac{\sum_{j=1}^{n_T} \sum_{h(j)=1}^{k} y_{h(j)} \times \omega_j^t(\mathbf{x})}{k\omega_1^t(\mathbf{x}) + k\omega_2^t(\mathbf{x}) + \ldots + k\omega_{n_T}^t(\mathbf{x})} \\
&= \frac{\sum_{j=1}^{n_T} y_j \times \omega_j^t(\mathbf{x})}{\sum_{j=1}^{n_T} \omega_j^t(\mathbf{x})} - \frac{\sum_{j=1}^{n_T} \sum_{h(j)=1}^{k} y_{h(j)} \times \omega_j^t(\mathbf{x})}{k \sum_{j=1}^{n_T} \omega_j^t(\mathbf{x})} \\
&= \sum_{j=1}^{n_T} \left[y_j \times \frac{\omega_j^t(\mathbf{x})}{\sum_{j=1}^{n_T} \omega_j^t(\mathbf{x})} \right] - \sum_{j=1}^{n_T} \sum_{h(j)=1}^{k} \left[y_{h(j)} \times \frac{\omega_j^t(\mathbf{x})}{k \sum_{j=1}^{n_T} \omega_j^t(\mathbf{x})} \right] \\
&= \sum_{j=1}^{n_T} y_j \times W_j^t - \sum_{j=1}^{n_T} \sum_{h(j)=1}^{k} y_{h(j)} \times W_j^c
\end{aligned}$$

Defining

$$\begin{aligned}
W_j^t &= \frac{\omega_j^t(\mathbf{x})}{\sum_{j=1}^{n_T} \omega_j^t(\mathbf{x})} \\
W_j^c &= \frac{\omega_j^t(\mathbf{x})}{k \sum_{j=1}^{n_T} \omega_j^t(\mathbf{x})}
\end{aligned}$$

We can conclude that

$$W_j^c = \frac{1}{k} W_j^t$$

Notice that each of the n_T treated units receives a weight of W_j^t and each of the comparison units a weight of $\frac{1}{k} W_j^t$.