Greg Ridgeway*, Stephanie Ann Kovalchik, Beth Ann Griffin and
Mohammed U. Kabeto

Propensity Score Analysis with Survey Weighted Data

DOI 10.1515/jci-2014-0039

Abstract: Propensity score analysis (PSA) is a common method for estimating treatment effects, but researchers dealing with data from survey designs are generally not properly accounting for the sampling weights in their analyses. Moreover, recommendations given in the few existing methodological articles on this subject are susceptible to bias. We show in this article through derivation, simulation, and a real data example that using sampling weights in the propensity score estimation stage and the outcome model stage results in an estimator that is robust to a variety of conditions that lead to bias for estimators currently recommended in the statistical literature. We highly recommend researchers use the more robust approach described here. This article provides much needed rigorous statistical guidance for researchers working with survey designs involving sampling weights and using PSAs.

Keywords: propensity score, sampling weights, survey weights

1 Introduction

Propensity score analysis (PSA) is now an exceedingly common statistical approach for estimating treatment effects from observational data. However, when the data are collected from a survey design, such as those that require the use of sampling weights, there is very limited guidance for researchers on how to use the sampling weights properly in their PSA. DuGoff et al. [1] searched for articles in health services research and found 28 studies using data containing design weights with an analysis involving propensity scores. Sixteen of those studies ignored the weights completely (with many nonetheless still claiming representativeness), seven used the weights only in the outcome model, and five used the weights in both the propensity score and outcome model. Unfortunately, we argue in this paper that only the latter group (the smallest group) adopted the best approach. Through derivation, simulation, and a real data example, we show that using sampling weights as observation weights in both the propensity score model and the outcome model provides robustness and failing to do so leaves analyses susceptible to bias.

Previous research has examined how the combination of sample selection and treatment selection can affect treatment effect estimation. Smith and Sugden [2] conclude that correct analyses depend on analysts considering “the joint distribution of the observations and of the sampling and assignment indicator variables.” While their analysis only briefly touches on propensity scores, their conclusion that a proper accounting of sampling design is essential still holds.

Two recent studies examined the question of survey design and PSA. Zanutto [3] writes “Since the propensity score model is used only to match treated and control units with similar background characteristics together in the sample and not to make inferences about the population-level propensity score model, it is not necessary to use survey-weighted estimation for the propensity score model.” While it may be true that we do not intend for the propensity score model to reflect a population, more precisely the choice to include or not to include sampling weights in the propensity score model hinges

***Corresponding author: Greg Ridgeway,** Department of Criminology, University of Pennsylvania, 3718 Locust Walk, Philadelphia, PA 19104-6286, USA, E-mail: gridge@sas.upenn.edu
Stephanie Ann Kovalchik, Beth Ann Griffin, RAND Corporation, Santa Monica, CA, USA
Mohammed U. Kabeto, Department of Internal Medicine, University of Michigan, Ann Arbor, MI, USA

on which approach produces better population-level treatment effect estimates. Zanutto [3] did not provide mathematical support for the recommendation to exclude sample weights from propensity score estimation.

DuGoff et al. [1] echo the assertion from Zanutto [3], “we argue that the propensity score model does not need to be survey-weighted, as we are not interested in generalizing the propensity score model to the population.” However, DuGoff et al. [1] do “recommend including the sampling weight as a predictor in the propensity score model” as they “may capture relevant factors, such as where individuals live, their demographic characteristics, and perhaps variables related to their probability of responding to the survey.” Still, the authors state that their guidance should not be the final say on how to properly implement PSAs with survey data as more work on the topic was needed.

The current recommendations in the literature conflicted with our intuition. We believe the majority of studies that involve survey designs are inherently designed to understand population-level effects and sampling weights should, therefore, be incorporated at every stage of estimation. Given the differing views on this topic, we set out to analyze more fully the proper use of sampling weights in PSA. Based on our investigations, we conclude that three scenarios in particular will cause problems if researchers use currently recommended approaches:

1. If there is a covariate z used in the sampling weights that is not used or available for the propensity score model even if z is independent of the potential outcomes.
2. If the propensity score model has limited degrees of freedom and spends those degrees of freedom on the domain of pretreatment covariates x with small sampling weights.
3. If the sampling probability depends on treatment assignment, particularly for the case when treatment and control cases are drawn from different survey efforts or different survey waves.

We find that using sampling weights in the propensity score estimation stage (as weights, not as a covariate), computing final weights as the product of the sampling weight and the propensity score weight, and using those final weights in an outcome model will be robust in these scenarios and be competitive in other considered scenarios.

In Section 2, we provide a theoretical justification for using the sampling weight (as a weight) when estimating propensity score weights and then highlight a number of cases where current methods proposed in the literature will be susceptible to bias in light of the theoretical derivation. In Section 3, we implement a simulation study that compares our proposed approach to standard methods often used in practice. In Section 4, we compare the methods using the Insights from the Newest Members of America’s Law Enforcement Community survey. Finally, in Section 5, we discuss the implications of our findings.

2 Theoretical justification for using sampling weights when estimating propensity score weights

Recent articles suggest that there is no need to use sampling weights in the propensity score model [1, 3]. To examine this assertion we need to work out our objective and connect it with the source of our sample. In this section, we focus on estimating the population average treatment effect on the treated (PATT), but the computations are analogous for other estimands and later sections with simulations and data analyses examine those other estimands.

We assume the standard data structure. We have a sample of n observations from a population of treated and untreated cases. For observation i we have t_i, the 0/1 treatment indicator, y_{1i}, the potential outcome if case i were assigned to treatment, y_{0i}, the control potential outcome, and $\mathbf{x}_i$, additional covariates possibly related to both t_i and (y_{1i}, y_{0i}). We observe either y_{1i} or y_{0i} depending on the value of t_i. In addition to this standard setup, we assume that observation i has been selected into the sample with probability p_i. As a result, the sample of treatment cases do not represent the population of treated cases nor does the sample of control cases represent the population of cases that were not assigned to treatment. However, assigning observation i a sampling weight of $1/p_i$ reweights the data so that they represent the population from which we drew them.

The expected treatment outcome for those assigned to treatment is straightforward and we can compute it directly from the data.

$$\mathrm{E}(y_1|t=1) \approx \frac{\sum_{i=1}^{n} t_i(1/p_i)y_{1i}}{\sum_{i=1}^{n} t_i(1/p_i)} \tag{1}$$

Note that eq. (1) is a design-based approach to estimation rather than a model-based approach. Little [4] provides an extensive review of the design-based and model-based approaches to estimation and their strengths and weaknesses. As in DuGoff et al. [1], in this paper we assume the analyst is pursuing a design-based approach and work through scenarios with that approach.

The challenge in causal analysis is to estimate the counterfactual $\mathrm{E}(y_0|t=1)$, the expected outcome of treatment cases if they had been assigned to the control condition. Standard propensity score approaches for estimating PATT, whether matching, stratifying, or weighting, all adjust the control cases so that they resemble the treatment cases on observed features. The complication that sampling weights add is that we need to rebalance the comparison cases so that the distribution of their features resembles the feature distribution of the treated *population*. How exactly to do that has generated confusion. While DuGoff et al. [1] aim "to make the observed treated and control groups as similar as possible," this approach will not yield a consistent treatment effect estimator at the population-level in many cases. If we make the distribution of the sampled control cases' features resemble the feature distribution of the treatment *population*, as in eq. (2), then we can achieve consistency. Mathematically stated, we want to find propensity score weights $w(\mathbf{x})$ such that

$$f(\mathbf{x}|t=1) = w(\mathbf{x})f(\mathbf{x}|t=0,s=1) \tag{2}$$

where $\mathbf{x}$ is a vector of case features, t is the 0/1 treatment indicator as before, and s is the 0/1 indicator of inclusion in the sample. Rearranging and applying Bayes Theorem twice we find

$$w(\mathbf{x}) = \frac{f(s=1,t=0)}{f(t=1)} \; \frac{1}{f(s=1|t=0,\mathbf{x})} \; \frac{f(t=1|\mathbf{x})}{1-f(t=1|\mathbf{x})} \tag{3}$$

These calculations show that the right weight for the comparison cases involve three terms. The first is a constant that will be absorbed into any normalization of $w(\mathbf{x})$. The second term is the standard inverse sampling probability weight. The third is the standard odds of treatment propensity score weight. However, those propensity score calculations are not conditional on $s=1$. Therefore, if $f(t=1|\mathbf{x}) \neq f(t=1|\mathbf{x},s=1)$ then $w(\mathbf{x})$ might not result in aligning the treatment and comparison groups to the right distribution of case features. Simply replacing $f(t=1|\mathbf{x})$ with an estimate based on $f(t=1|\mathbf{x},s=1)$ risks not satisfying the primary goal of propensity scoring, aligning the right feature distributions.

The next several subsections explore how various scenarios and estimation methods affect covariate balance and treatment effect estimates.

2.1 Effect of missing sampling weight variables when estimating the propensity score model

In this subsection, we introduce a key complication to demonstrate the risks associated with estimators that do not include sampling weights in the propensity score estimation step. Namely, we introduce a variable z that is used in the construction of the sampling weights, but is unavailable when estimating the propensity score. Such zs are not so unusual, particularly for datasets that are part of the federal statistical system. For example, sampling weights often include features related to fielding method, geography, and household, any of which could be related to outcomes of interest. These features might not be available (or might not be available at the same resolution) to the analyst. There is also the risk that the analyst might not be entirely aware of what features were used in constructing the design weights even if all of those features are technically available.

Combining eqs (1) and (3), the estimator of PATT has the form

$$\widehat{\mathrm{PATT}} = \frac{\sum_{i=1}^{n} y_i t_i (1/p_i)}{\sum_{i=1}^{n} t_i (1/p_i)} - \frac{\sum_{i=1}^{n} y_i (1-t_i)(e_i/1-e_i)(1/p_i)}{\sum_{i=1}^{n} (1-t_i)(e_i/1-e_i)(1/p_i)} \tag{4}$$

where $e_i = f(t=1|\mathbf{x}_i)$, the propensity score. The first term estimates $\mathrm{E}(y_1|t=1)$ regardless of the relationship of z to the treatment assignment or to the potential outcomes. We write out the details of this well-known property to more easily guide the analysis into the more complicated second term.

$$\frac{\sum_{i=1}^{n} y_i t_i (1/p_i)}{\sum_{i=1}^{n} t_i (1/p_i)} \to \frac{\iiint y_1 \frac{1}{f(s=1|y_1,\mathbf{x},z,t=1)} f(y_1,\mathbf{x},z|s=1,t=1)\, dy_1\, d\mathbf{x}\, dz}{\iiint \frac{1}{f(s=1|y_1,\mathbf{x},z,t=1)} f(y_1,\mathbf{x},z|s=1,t=1)\, dy_1\, d\mathbf{x}\, dz} \tag{5}$$

$$= \frac{\iiint y_1 \frac{f(s=1|y_1,\mathbf{x},z,t=1) f(y_1,\mathbf{x},z|t=1)}{f(s=1|y_1,\mathbf{x},z,t=1) f(s=1|t=1)}\, dy_1\, d\mathbf{x}\, dz}{\iiint \frac{f(s=1|y_1,\mathbf{x},z,t=1) f(y_1,\mathbf{x},z|t=1)}{f(s=1|y_1,\mathbf{x},z,t=1) f(s=1|t=1)}\, dy_1\, d\mathbf{x}\, dz} \tag{6}$$

$$= \frac{\frac{1}{f(s=1|t=1)} \iiint y_1 f(y_1,\mathbf{x},z|t=1)\, dy_1\, d\mathbf{x}\, dz}{\frac{1}{f(s=1|t=1)} \iiint f(y_1,\mathbf{x},z|t=1)\, dy_1\, d\mathbf{x}\, dz} \tag{7}$$

$$= \mathrm{E}(y_1|t=1) \tag{8}$$

The second term in eq. (4) should estimate $\mathrm{E}(y_0|t=1)$, but we will see that this depends on how we estimate the propensity scores, p_i, and on a few key assumptions. Here we will just analyze the numerator since, as seen in eq. (7) the denominator is simply a normalization term.

If we use the approach described in Zanutto [3] and DuGoff et al. [1] then we insert the propensity score conditional on $s=1$ in place of e_i.

$$\sum_{i=1}^{n} y_i (1-t_i) \frac{e_i}{1-e_i} \frac{1}{p_i} \to n \iiint y_0 \frac{f(t=1|\mathbf{x},s=1)}{f(t=0|\mathbf{x},s=1)} \frac{f(y_0,\mathbf{x},z|t=0,s=1)}{f(s=1|y_0,\mathbf{x},z,t=0)}\, dy_0\, d\mathbf{x}\, dz \tag{9}$$

$$= n \iiint y_0 \frac{f(t=1|\mathbf{x},s=1,y_0)}{f(t=0|\mathbf{x},s=1,y_0)} \frac{f(y_0,\mathbf{x},z|t=0)}{f(y_0,\mathbf{x},z|t=0,s=1) f(s=1|t=0)} f(y_0,\mathbf{x},z|t=0,s=1)\, dy_0\, d\mathbf{x}\, dz \tag{10}$$

In eq. (10) we need t to be independent of y_0 conditional on $\mathbf{x}$, the standard independence assumption used in PSA, in order to insert y_0 in the treatment conditional probabilities. Integrating out z we obtain

$$= \frac{n}{f(s=1|t=0)} \iint y_0 \frac{f(t=1|y_0,\mathbf{x},s=1)}{f(t=0|y_0,\mathbf{x},s=1)} f(y_0,\mathbf{x}|t=0)\, dy_0\, d\mathbf{x} \tag{11}$$

$$= \frac{n}{f(s=1|t=0)} \frac{f(t=0)}{f(t=1)} \iint y_0 \frac{f(y_0,\mathbf{x}|t=0)}{f(y_0,\mathbf{x},s=1|t=0)} f(y_0,\mathbf{x},s=1|t=1)\, dy_0\, d\mathbf{x} \tag{12}$$

Bringing in the denominator normalizing term we arrive at

$$\frac{\sum_{i=1}^{n} y_i (1-t_i)(e_i/(1-e_i))(1/p_i)}{\sum_{i=1}^{n} (1-t_i)(e_i/(1-e_i))(1/p_i)} \to \iint y_0 \frac{f(y_0,\mathbf{x}|t=0)}{f(y_0,\mathbf{x},s=1|t=0)} f(y_0,\mathbf{x},s=1|t=1)\, dy_0\, d\mathbf{x} \tag{13}$$

However, eq. (13) shows that the estimator is not necessarily consistent for $\mathrm{E}(y_0|t=1)$. We need $f(y_0,\mathbf{x},s|t) = f(y_0,\mathbf{x}|t) f(s|t)$ but the independence of s and $(y_0,\mathbf{x})$ cannot be guaranteed when z, even though no longer directly visible in the expression, might have induced a correlation.

To repair the treatment effect estimator, we need to modify eq. (9) by using the sampling weights in the estimation of the propensity score model so that we obtain a consistent estimate of $f(t=1|\mathbf{x})$. Doing so results in

$$\sum_{i=1}^{n} y_i(1-t_i)\frac{e_i}{1-e_i}\frac{1}{p_i} \to n\iiint y_0 \frac{f(t=1|\mathbf{x})}{f(t=0|\mathbf{x})}\frac{f(y_0,\mathbf{x},z|t=0,s=1)}{f(s=1|y_0,\mathbf{x},z,t=0)}\,dy_0\,d\mathbf{x}\,dz \tag{14}$$

$$= \iint y_0 f(y_0,\mathbf{x}|t=1)\,dy_0\,d\mathbf{x} \tag{15}$$

$$= \mathrm{E}(y_0|t=1) \tag{16}$$

The following simple example shown in Table 1 makes the situation more concrete. In this example, the sampling and treatment assignment probabilities depend on z, but the potential outcomes do not depend on z.

Table 1: Example scenario with sampling and treatment probabilities dependent on z.

N	x	z	$P(s\|x,z)$	$P(t=1\|x,z)$	$\mathrm{E}(y_0\|x,z)$	$\mathrm{E}(y_1\|x,z)$
1,000	0	0	0.2	0.1	1	1
1,000	0	1	0.3	0.9	1	1
1,000	1	0	0.4	0.8	4	4
1,000	1	1	0.5	0.8	4	4

The inclusion or exclusion of the sampling weights in the propensity score model and the necessary assumptions do in fact matter. Table 2 compares the asymptotic results for this example. The first column in Table 2 shows the asymptotic mean of x and y for the treated group. The second column shows using sampling weights in both stages of the PSA results in balance on x and an $\mathrm{E}(y_0|t=1)$ identical to the treatment group, consistent with the actual null treatment effect simulated here. The third column shows that without sampling weights in the propensity score model we do not get balance on x and do not correctly estimate the null treatment effect.

Table 2: Asymptotic values for methods with and without sampling weights in the propensity score model.

	$t=1$	$t=0$	$t=0$
Sampling weight PS model		Yes	No
Sampling weight outcomes	Yes	Yes	Yes
$\mathrm{E}(x\|t)$	0.615	0.615	0.537
$\mathrm{E}(y_t\|t=1)$	2.846	2.846	2.610

This analysis demonstrates that when there are case features used in the development of the sampling weights that are unavailable when estimating the propensity score, consistent estimates depend on additional independence assumptions, assumptions that are not needed if the propensity score estimator uses the sampling weights.

2.2 Propensity score models with limited degrees of freedom

Even if the situation described in the previous example does not occur, the propensity score estimates should still involve the sampling weights. While we recommend the use of readily available non-parametric propensity score estimators [5–7], we recognize that propensity score models with limited degrees of freedom, such as standard logistic regression models, are exceedingly common. Ideally, researchers will use propensity score methods that are flexible, can accommodate non-linearities and interactions, and are not sensitive to outliers and high leverage cases. However, in practice we find researchers not taking advantage of more flexible methods.

Analysts using such parametric models should focus the expenditure of their limited degrees of freedom on the domain of **x** with the largest weights. Equation (3) shows that the final analytical weight on a control case will be the case's sampling weight times the propensity score weight,

$$\frac{1}{p_i} \times \frac{f(t=1|\mathbf{x}_i)}{1-f(t=1|\mathbf{x}_i)}.$$

Clearly the quality of the propensity score will matter more for values of **x** where $1/p_i$ is large and will be inconsequential for cases with $1/p_i$ near 0. Therefore, the propensity score model should allocate available degrees of freedom to the regions of **x** with the largest weight.

The example shown in Figure 1 demonstrates this. In this example the treatment probability has a curvilinear relationship with x (black). Because in this example we limit ourselves to a propensity score model with two degrees of freedom, the resulting propensity score model will vary based on which cases have the most weight.

The density plots shown in the top margin show that $f(x)$ (blue), which puts more mass on larger values of x, is quite different than $f(x|s=1)$ (red), generated by uniformly sampling across x. The red curve overlaying the treatment probability is the result of a standard logistic regression model with an intercept and a coefficient for x. The limited capacity of such a model to learn the treatment's relationship with x means that it results in a mediocre fit everywhere along x. The blue curve is the result of a logistic regression model fit using inverse sampling probability weights. The result is that this propensity score model produces a poor fit for $x < 40$, but a good fit for $x > 40$, the region with the highest sampling weights.

The weighted control mean of x when using the sampling weights in the propensity score is 65.67, quite close to the treatment mean of 65.45. Using a propensity score model without sampling weights the control group mean is 63.48. If x were a feature like age, then a two year difference between the treatment and

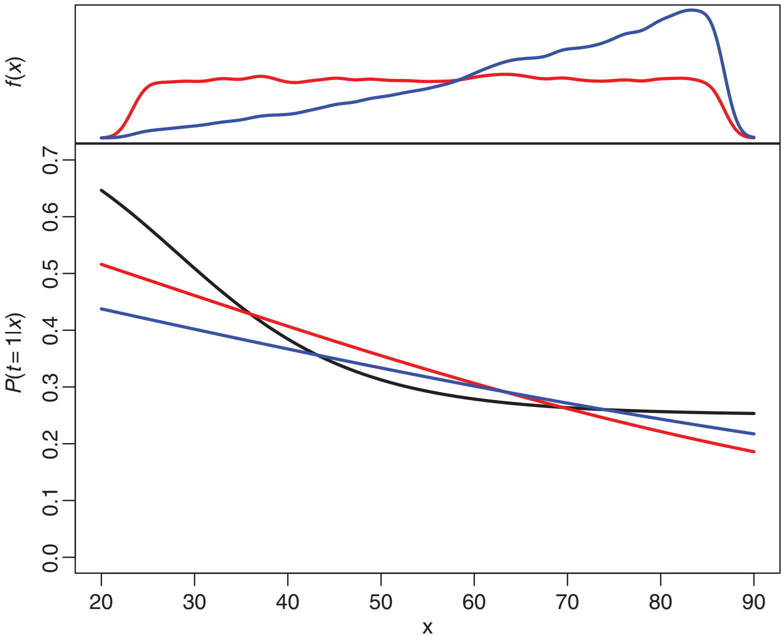

Figure 1: Example propensity score models. The black curve traces the true treatment assignment probability. The red curve is the estimated propensity score without sampling weights. The blue curve is the propensity score estimated using the weights. The plot in the top margin shows the distribution of x in the sample (red and nearly uniform) and the distribution of x weighted to reflect the population (blue and increasing with x).

control group reasonably could be a confounder for morbidity and mortality outcomes. A consequence of this would be that it will be harder to find good balance if an analyst is using the wrong (unweighted) propensity score model and easier when using the weighted one.

Using the sampling weights as described in this subsection can reduce the effect of model misspecification of some kinds, but is not a cure-all. This example involved modest non-linearity, most of which was in a region with low sampling weight and so using the sampling weights produced a better propensity score model. Cases with large sampling weight and high leverage (i.e. outliers with large sampling weights) will cause estimation problems for the propensity score model with or without weights, and possibly for the outcome analysis as well. This problem could be more completely resolved by directly addressing propensity score model misspecification. However, the example shown here demonstrates that modest misspecification can be accommodated when using sampling weights.

2.3 Issues when sampling weights drawn from different sources

Additional complications arise for propensity score estimators that are not weighted by the sampling weights when the treatment group and the control group come from different survey efforts. For example, data fusion is a form of statistical data linkage that does not aim to match individual records into two data sources, but rather match a collection of cases in the two data sources that have similar features. Such questions arise in comparisons between military spouses and similar members of the general public [8], comparing television viewing and consumer behavior [9], and when comparing any study sample to government or administrative surveys with similar measures. In these cases, the data needed to answer the question reside in two different data collections and the sampling weights for one dataset are not normalized to the same scale as the comparison data source.

Respondents with the same sampling weight from different surveys will not share the same features. Using the sampling weight as a covariate (as suggested in DuGoff et al. [1]) in such circumstances will result in bias. Using the sampling weights as weights does not have the same issue. The sampling weights are still derived from valid sample inclusion probabilities. Therefore, the PATT estimator defined in eq. (4) is still valid. Any rescaling of the weights in either sample will have no effect as the scaling factor will cancel out. Regarding the propensity score model, only the intercept term will be affected by a rescaling of the weights and that too cancels out of eq. (4). When estimating the population average treatment effect (PATE), the propensity score model intercept term does not cancel out of the estimator. Therefore, when estimating PATE the control cases' weights should be scaled so that their share of the total weight matches the fraction of control cases in the population of interest.

This is a case in which the sampling probability depends on the treatment assignment. As we saw in Section 2.1, correlations between sampling probabilities and treatment assignment (in that case possibly induced by z) can produce inconsistent estimators. Rather than conduct another derivation showing the special issue that a dependence between sampling probability and treatment creates, we include this scenario in our simulated examples in the following section.

3 Simulation study

DuGoff et al. [1] previously reported results of a simulation study comparing the bias and coverage of different estimators of the PATE and PATT using survey data. In this section, we expand on their study in order to examine the comparative performance of a broader set of propensity score estimators and modeling scenarios. The simulation is the same as DuGoff et al. [1]: a single normally distributed covariate x, with mean conditional on the population stratum; a binary treatment indicator t; normally distributed potential outcomes y_0, y_1 with a heterogeneous treatment effect that varies with x; and a stratified population of

90,000 persons with three strata of equal size but unequal probabilities of selection. Details of the simulation are included in Appendix A1.

The five data generation scenarios investigated in our simulation study are listed across the top of Tables 3 and 4. For each scenario, we vary the selection probability model to include cases where selection depends on x ($s \sim x$), selection is independent of x ($s \perp x$), selection depends on both x and t ($s \sim (x, t)$), and where selection depends on x with weights differentially scaled by t to simulate samples generated from different survey efforts ($s \sim x|t$). We also varied the treatment probability model to depend on x ($t \sim x$) or to have a non-linear relationship with x ($t \sim x^2$).

For each of the five data generation scenarios, we examined the performance of four candidate propensity score approaches, listed down the side of Table 3, including no PSA (None, $t \sim 1$), estimating the propensity score ignoring the sampling weights (None, $t \sim x$), estimating the propensity score with the sampling weight as a covariate (Covariate, $t \sim x + sw$), and estimating the propensity score using the sampling weight as an observation weight (Weight, $t \sim x$). PATE estimators use propensity score weights equal to $1/\hat{f}(t = 1|x)$ for the treated and $1/\hat{f}(t = 0|x)$ for controls. PATT estimators use propensity score weights equal to 1 for all treated and $\hat{f}(t = 1|x)/\hat{f}(t = 0|x)$ for controls. Our study focuses on population treatment effects so we did not examine within-sample treatment effects.

We first judged the performance of the estimators by examining the balance of the covariate x after propensity score weighting. We measured covariate balance using the population standardized mean difference (SMD), equal to the mean difference in x between the treated and control group, weighted by the product of the propensity score weight and sampling weight, and divided by the pooled standard deviation of x in the survey sample.

In Table 3 we compare the balance achieved by each of the estimators. Only propensity score models using sampling weights as weights consistently provided good covariate balance across the scenarios. In contrast all other methods had at least one scenario that resulted in poor covariate balance. This was particularly true of the method using the sampling weight as a covariate. These findings illustrate the potential problems with using sampling weights as a covariate when the sampling weights are associated with treatment group (Scenario 3); when the treatment groups being compared are from different target populations (Scenario 4),

Table 3: Standardized mean differences for measuring covariate balance.

			Data generation scenario				
			1	**2**	**3**	**4**	**5**
		Description	**Standard**	**Random sample**	**Selection depends on *x* and *t***	**Weight scales differ**	**Non-linear treatment**
		Selection	$s \sim x$	$s \perp x$	$s \sim (x, t)$	$s \sim x\|t$	$s \sim x$
		Treatment	$t \sim x$	$t \sim x$	$t \sim x$	$t \sim x$	$t \sim x^2$
	Fit PS model						
	Sampling weights	**PS model**			**PATE**		
1	None	$t \sim 1$	1.18	1.06	1.10	1.17	0.13
2	None	$t \sim x$	0.10	0.03	0.03	0.09	2.00
3	Covariate	$t \sim x + sw$	0.06	0.03	2.28	1.39	0.89
4	Weight	$t \sim x$	0.04	0.03	0.03	0.04	0.01
					PATT		
5	None	$t \sim 1$	1.18	1.06	1.10	1.17	0.13
6	None	$t \sim x$	0.17	0.03	0.08	0.16	0.74
7	Covariate	$t \sim x + sw$	0.12	0.03	1.57	0.44	1.10
8	Weight	$t \sim x$	0.12	0.02	0.06	0.11	0.08

Notes: Covariate balance measured with absolute standardized mean differences. All models fit use sampling weights when computing these differences, but vary on whether or how those sampling weights were used in the propensity score model.

Table 4: Root mean squared error of population treatment effect estimates.

				Data generation scenario				
				1	2	3	4	5
			Description	**Standard**	**Random sample**	**Selection depends on *x* and *t***	**Weight scales differ**	**Non-linear treatment**
			Selection	$s\sim x$	$s\perp x$	$s\sim(x,t)$	$s\sim x\mid t$	$s\sim x$
			Treatment	$t\sim x$	$t\sim x$	$t\sim x$	$t\sim x$	$t\sim x^2$
		Fit model						
	Sampling weights	**PS model**	**Outcome model**			**PATE**		
1	None	$t\sim 1$	$y\sim t+x$	0.051	0.037	0.043	0.064	0.032
2	None	$t\sim x$	$y\sim t+x$	0.034	0.023	0.028	0.036	0.138
3	Covariate	$t\sim x+sw$	$y\sim t+x$	0.035	0.023	4.075	0.085	0.459
4	Weight	$t\sim x$	$y\sim t+x$	0.034	0.023	0.026	0.033	0.032
5	None	$t\sim 1$	$y\sim t$	1.111	1.150	1.125	1.116	0.160
6	None	$t\sim x$	$y\sim t$	0.142	0.045	0.046	0.125	1.957
7	Covariate	$t\sim x+sw$	$y\sim t$	0.093	0.045	2.273	1.685	1.031
8	Weight	$t\sim x$	$y\sim t$	0.058	0.045	0.050	0.055	0.035
						PATT		
9	None	$t\sim 1$	$y\sim t+x$	0.049	0.041	0.045	0.042	0.032
10	None	$t\sim x$	$y\sim t+x$	0.066	0.021	0.037	0.069	0.045
11	Covariate	$t\sim x+sw$	$y\sim t+x$	0.069	0.021	3.486	0.042	0.621
12	Weight	$t\sim x$	$y\sim t+x$	0.066	0.021	0.037	0.069	0.035
13	None	$t\sim 1$	$y\sim t$	1.041	1.079	1.055	1.046	0.160
14	None	$t\sim x$	$y\sim t$	0.230	0.041	0.109	0.214	0.674
15	Covariate	$y\sim t+x$	$y\sim t$	0.161	0.041	1.615	0.401	1.175
16	Weight	$t\sim x$	$y\sim t$	0.155	0.039	0.093	0.149	0.113

Notes: (1) All methods use sampling weights in the outcome model, but vary on whether and how those sampling weights were used in the propensity score model. (2) We ran 2,000 simulations, sufficient to estimate the RMSE for any competitive estimator to within ±0.001.

as described in Section 2.3; or when the propensity score model is misspecified (Scenario 5). When sampling weights were strongly associated with treatment assignments, the propensity scores would perfectly separate treated and control cases. We found a similar set of issues for estimators of PATT.

We also evaluated the candidate treatment effect estimators in terms of their root mean squared error (RMSE) when estimating the true population treatment effect. Table 4 again lists the five data generation scenarios along the top and the four propensity score methods down the side. We also considered two different outcome models. The first includes x as a covariate ($y\sim t+x$). This is the approach used in DuGoff et al. [1]. While this is an appropriate model and provides a doubly robust treatment effect estimate, this is not the typical propensity score approach. Therefore, we also include the more typical PSA estimator that excludes x in the outcome model ($y\sim t$). In this simulation x is a strong predictor of the outcome and, therefore, outcome models that include x will generally have smaller RMSE. All outcome models used weights equal to the product of the sampling weight and the propensity score weight.

We found that the estimator using the sampling weights as observation weights in the development of the propensity score model had among the smallest RMSE of all the estimators across the range of scenarios. Particularly in contrast to using the propensity score that incorporates sampling weights as a covariate, using sampling weights at all stages of PSA produces notably more precise treatment effect estimates for scenarios 3 through 5. These findings mirror the covariate balance findings.

We also examined the coverage and Type I error of the population estimators presented in this section. Coverage describes the frequency with which the true population treatment effect was captured by a 95% confidence interval for the given estimator, which depends on bias and standard error properties of the estimator. Type I error refers to the frequency that the coverage interval does not include zero in scenarios in which there was no treatment effect. For the set of scenarios considered in this study, no important differences in coverage or Type I error were found among the different approaches for estimating population treatment effects.

We also checked to make sure that the findings shown here were not the result of common support issues. The data generating process used in the simulation study specified a common support for both treatment groups. That is, the theoretical range of x was equal for treated and controls. However, any finite sample might result in extreme cases in each treatment group, such that, no comparable "match" can be found. The proportion of these cases was small. For example, if we were to have truncated the samples to the observed common support based on the estimated propensity score, only 1.5% of the sample would have been discarded in Scenario 1 and only 2% in Scenario 5, on average. Importantly, there was no difference in the common support depending on the method of propensity score estimation used (e.g. no survey weight, with survey weight as covariate, or weighted by survey weight). This indicates that the comparative performance evaluation was not sensitive to exclusions based on the common support.

In summary, this simulation study demonstrates that most propensity-weighted estimators of population treatment effects perform well under ideal conditions, in which the survey design is not complicated and the propensity score model and outcome model are correctly specified. However, when the study design becomes more complex or model misspecification is present, a sample-weighted propensity score model reduces population covariate imbalances and produces more robust and more accurate causal effect estimates. Given these are scenarios we believe are most likely to occur in real applications, the safest bet in any particular analysis will be to use the weights at all stages of an analysis.

4 Insights from the newest members of America's law enforcement community

In this section, we show that the inclusion or exclusion of sampling weights in the propensity score model stage can affect covariate balance and study outcomes in a real dataset.

For much of the last 15 years recruiting has been among the greatest challenges for the law enforcement community, particularly for large municipal agencies seeking to develop a larger, diverse workforce well suited to community-oriented policing. The Office of Community Oriented Policing Services (COPS) in the U.S. Department of Justice was interested in learning more about the barriers new and potential recruits face when considering a law enforcement career. The 2009 Insights from the Newest Members of America's Law Enforcement Community survey [10] was conducted to aid the law enforcement community in refining its recruitment practices and improving recruitment results. The survey targeted new law enforcement recruits, reaching a national pool of 1,600 respondents from 44 of the United States' largest police and sheriff departments. The survey asked recruits about their reasons for pursuing a career in law enforcement, potential disadvantages of such a career, influencers on a career in law enforcement and employment within the recruit's chosen agency, and the perceived effectiveness of both actual and potential recruiting strategies.

One research question of interest in the study included understanding how attractions and barriers to joining a law enforcement agency differed for minority and white respondents. Recruiting minority officers is particularly challenging, both at the time of the fielding of the survey and today following nationwide protest concerning police use of lethal force against unarmed minorities. However, confounding is likely since minority recruits were more likely to be female, be married, have children, and have never attended college than white recruits. Moreover, minority respondents in the survey were a select sample who needed to be reweighted using the sampling weights in the study to ensure they represented the general population of new minority recruits. Respondents were more likely to be married, more likely to have children, and less likely to have attended college. Therefore, correct analyses depend on proper use of propensity scores and sampling weights.

Propensity score analyses are useful in such circumstances to untangle the effects of race from these other respondent features. We considered four approaches for adjusting for covariance imbalances between minority and white recruits. The four methods varied on propensity score method (no propensity score adjustment, propensity score estimated without sampling weights, and propensity score estimated with sampling weights) and whether sampling weights were used when computing the SMD for each covariate (without sampling weights and with sampling weights).

Figure 2 shows the SMD for each method for 22 covariates, which included demographic, educational, and work history factors. Each line in the plot traces the path of one of 22 covariates as we used different survey adjustment methods. From the left side of Figure 2 we see that the raw data show large SMDs between minority and white respondents with many covariates having SMDs in excessive of 0.2. The SMDs shown on the far right of Figure 2 are from the method that the previous sections of this paper support, using sampling weights at all stages of analysis. The SMDs are all within 0.1 of zero. The two other approaches shown in the middle, using no propensity scores or fitting propensity scores without using sampling weights, both fare worse in terms of covariate balance.

The extent to which achieving better covariate balance has an impact on the results depends on the strength of the relationship between those covariates and the outcomes of interest. In this example, the choice of method impacts some conclusions from the survey, primarily in terms of the magnitude of effects. For example, results from the survey show that minority officers are significantly more concerned about benefits, particularly health insurance. If we do not use weights in the propensity score model, the analysis yields an odds ratio of 2.68 with a 95% confidence interval of (0.97, 7.41). Using the sample weights in both the propensity score model and the outcomes analysis yields and odds ratio of 2.56 (1.00, 6.53), a more precise estimate and, for those who are particular about an $\alpha = 0.05$ confidence level, a 95% confidence interval that does not overlap 1.0. The survey also suggests that minority officers are more concerned about police excessive force to the point that

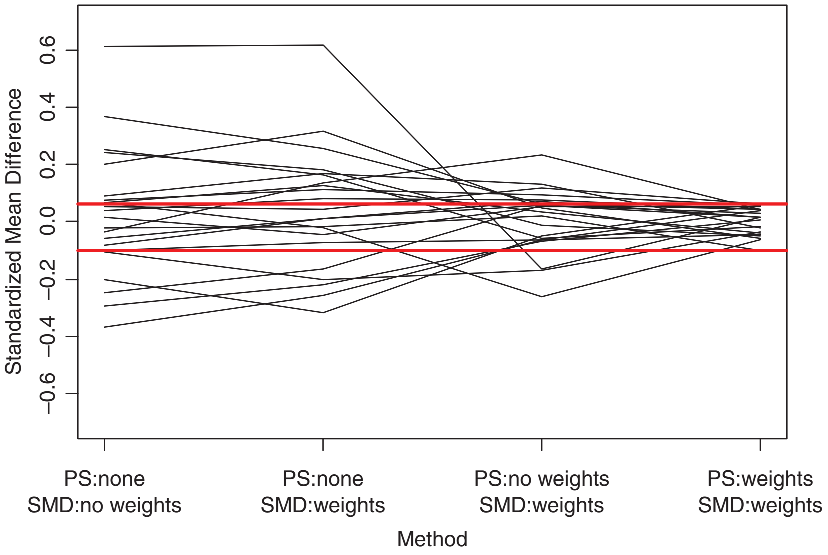

Figure 2: Population covariate balance contrasting the population standardized mean differences (SMD) between minority and white new police recruits for 22 individual-level covariates. We calculate SMD for each of the 22 covariates using four different methods, each using propensity score weights and sampling weights in different ways. Each line in the plot tracks one of the 22 covariates across the four methods. The red horizontal lines mark the range of SMD for the method using sampling weights at all stages of the modeling. Values near zero represent better balance.

they considered not joining. Without sampling weights in the propensity score model the estimated odds ratio is 1.64 (0.61, 4.44), but with sampling weights in the propensity score model the estimated odds ratio is 1.93 (0.92, 4.06). While neither estimate reaches standard levels of statistical significance, the first estimate likely would make the analyst dismiss the issue altogether while the second would draw some attention.

5 Discussion

PSA is a widely used technique and DuGoff et al. [1] show that many researchers are struggling to properly incorporate complex survey designs into their PSA. The analysis presented here can help researchers gain a better understanding of how best to handle sampling weights in PSA. In this paper, we showed that estimating propensity score models without the sampling weights has risks and will not be consistent in some cases:

1. If the sampling weights involve variables that are unavailable to the analysts estimating the propensity score model even if the missing variables are unrelated to the potential outcomes.
2. If degrees of freedom of the propensity score model are wasted on regions of $\mathbf{x}$ with small sampling weight resulting in poor model fit in the regions of $\mathbf{x}$ that matter most.
3. If the sampling probability depends on treatment assignment, particularly for the case when treatment and control cases are drawn from different survey efforts or different survey waves.

Through simulation we showed that across a range of scenarios the most robust strategy is to use the sampling weights in the propensity score model and to use the sampling weight times the propensity score weight as the weight in the final outcome analysis. Whether the use of sampling weights in the analysis will affect the conclusions depends on the strength of the relationships between outcomes, treatment, covariates, and sampling weights. We found that for some outcomes for the Insights from the Newest Members of America’s Law Enforcement Community survey the choice of analytical method mattered. The choice of method had a great effect on the quality of the covariate balance and a modest effect on conclusions drawn from the results.

We note that our findings here generalize readily to best practices for incorporating any kind of survey weight, such as nonresponse weights. If a study has computed nonresponse weights to make the sample of responders representative of the original baseline sample, our analysis here indicates that the use of the nonresponse weight (as a weight and not a covariate) in the propensity score model will lead to better inferences. This article focuses on simple survey designs which only entail sampling weights. However, we believe similar logic will apply for more complex survey designs and future research should carefully explore the impact that more complex surveys designs have on PSAs. Future work should also explore comparisons of the design-based approaches to model-based approaches in cases where differences in survey weights will be large between the treatment and control groups and the increase in variance from the design-based approach might lead to a preference for a model-based approach.

Acknowledgments: The authors are grateful to Dr. Elizabeth Stuart and Dr. Dan McCaffrey for their thoughtful feedback on our work and to Dr. Nelson Lim for prompting us to address this question.

Funding: Research reported in this article was supported by NIDA under award number R01DA034065 for Griffin’s work and by the NIA under award number 1R01AG043960 for Kabeto’s work. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Appendix

A.1 Details of the simulation

This appendix describes the simulation study in more detail. The R script for the simulation is available at JCI webpage for supplementary material.

First, we generate a simulated population of 90,000 observations divided into three strata of 30,000 cases each. Each observation has a single covariate $X \sim N(\mu_j, 1)$ where μ_j varies by the three strata, $\mu_j = \frac{1}{4} j - \frac{1}{2}$ for $j = 1, 2, 3$.

Treatment probabilities are $\text{logit}\, P(t=1|x) = -1 + \log(4)x$ except for Scenario 5 where $\text{logit}\, P(t=1|x) = -1 + \log(4)x^2$.

Selection probabilities are $\text{logit}\, P(s=1|x,t) = -2.8 - \log(4)x$ except for Scenario 2 where $\text{logit}\, P(s=1|x,t) = -2.8$ and Scenario 3 where

$$\text{logit}\, P(s=1|x,t) = -2.8 - \frac{1}{2}\log(4)x - \frac{1}{2}\log(4)t.$$

In Scenario 4 we leave the selection probabilities unchanged by artificially scaling the sampling weights of control cases by 1.3 to simulate the scenario of treatment and control case coming from different data collections with mismatched weight scales.

We simulate potential outcomes as $y_0 \sim N(1 + x, \frac{1}{4})$ and $y_1 \sim N(y_0 + 0.2 + 0.1x, \frac{1}{4})$.

All performance measures were based on summaries of 2,000 iterations of the indicated scenario.

References

1. DuGoff EH, Schuler M, Stuart EA. Generalizing observational study results: applying propensity score methods to complex surveys. Health Serv Res 2014;49:284–303.
2. Smith TMF, Sugden RA. Sampling and assignment mechanisms in experiments, surveys and observational studies. Int Stat Rev 1988;56:165–80.
3. Zanutto EL. A comparison of propensity score and linear regression analysis of complex survey data. J Data Sci 2006; 4:67–91.
4. Little RJ. To model or not to model? Competing modes of inference for finite population sampling. J Am Stat Assoc 2004;99:546–56.
5. Lee BK, Lessler J, Stuart EA. Improving propensity score weighting using machine learning. Stat Med 2010;29:337–46.
6. McCaffrey D, Ridgeway G, Morral A. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. Psychol Methods 2004;9:403–25.
7. McCaffrey DF, Griffin BA, Almirall D, Slaughter ME, Ramchand R, Burgette LF. A tutorial on propensity score estimation for multiple treatments using generalized boosted models. Stat Med 2013;32:3388–414.
8. Harrell MC, Lim N, Castaneda LW, Golinelli D. Working around the military: Challenges to military spouse employment and education. Technical Report MG-196-OSD, RAND Corporation, Santa Monica, CA, 2004. Accessed December 15, 2014. Available at: http://www.rand.org/pubs/monographs/MG196.html.
9. Rässler S. Statistical matching: a frequentist theory, practical applications and alternative Bayesian approaches, Lecture Notes in Statistics. New York: Springer, 2002.
10. Castaneda LW, Ridgeway G. Today's police and sheriff recruits: insights from the newest members of America's law enforcement community, Technical Report MG-992-DOJ, RAND Corporation, Santa Monica, CA, 2010. Accessed December 15, 2014. Available at: http://www.rand.org/pubs/monographs/MG992.html.