Efficient Covariate Balancing for the Local Average Treatment Effect

Phillip Heiler

Department of Economics and Business Economics, Aarhus University, CREATES, TrygFonden's Centre for Child Research, Denmark

ABSTRACT

This article develops an empirical balancing approach for the estimation of treatment effects under two-sided noncompliance using a binary instrumental variable. The method weights both treatment and outcome information with inverse probabilities to impose exact finite sample balance across instrument level groups. It is free of functional form assumptions on the outcome or the treatment selection step. By tailoring the loss function for the instrument propensity scores, the resulting treatment effect estimates are automatically weight normalized and exhibit both low bias and reduced variance in finite samples compared to conventional inverse probability weighting methods. We provide conditions for asymptotic normality and semiparametric efficiency and demonstrate how to use additional information about the treatment selection step for bias reduction in finite samples. A doubly robust extension is proposed as well. Monte Carlo simulations suggest that the theoretical advantages translate well to finite samples. The method is illustrated in an empirical example.

1. Introduction

Estimation of causal effects is at the heart of modern empirical research in economics. In particular, evaluating the impact of policies or programs on units with heterogeneous preferences and characteristics such as households, workers, unemployed, firms, or students is necessary to develop a thorough understanding of fundamental economic relationships. This article deals with the problem of estimating a causal effect of a treatment using variation from a binary independent instrumental variable (IV) when units have heterogeneous responses to the instrument and to the treatment. We develop a semiparametric estimation method for the local average treatment effect (LATE) based on inverse probability weighting (IPW). The method is designed to increase internal validity by reducing the finite sample bias in estimation of the treatment effect while allowing for dependent observable and unobservable variables to affect both treatment participation and potential outcomes of interest (selection on unobservables). It is more robust than conventional instrumental variable methods as it does not require parametric functional form assumptions about outcome or treatment selection steps or other restrictions such as constant causal effects. Moreover, it has appealing point estimation properties compared to conventional IPW estimators for the LATE.

The key insight required is that IPW estimation of the LATE fundamentally rests on balancing the distribution of observable confounders for two reduced form type components. A reduced form type component or effect refers to what corresponds to the slope parameter of a regression on the instrument in conventional IV analysis. In particular, for the reduced form type component of the outcome, observations that are instrumented at a given level are weighted by their inverse probability of being in that state, that is, by their instrument propensity scores. This assures that differences in observed confounders are correctly taken into account when estimating the reduced form type effect for units that choose treatment in accordance with the instrument (compliers). For the overall LATE, the remaining differences due to heterogeneous treatment selection responses to the instrument are incorporated by a second reduced form type component that applies the same inverse probability weights to the corresponding treatment indicators.

On a technical level, balancing means that the averages of the inverse probability weighted observable covariates between groups defined by the instrument are imposed to be identical. However, conventional methods for estimation of the weights such as maximum likelihood or even true weights do not yield perfect balance in finite samples. We exploit recent advances in the literature on the evaluation of causal effects established for conditionally independent treatments (selection on observables) to construct empirical balancing conditions for the estimation of inverse probability weights. Finite sample balance is achieved through tailoring the loss function for the instrument propensity scores used in the estimation of the LATE. The method compares favorably to conventional IPW methods as the tailored loss approach minimizes approximate bias while simultaneously favoring weights that do not exhibit too much variance. In addition, it preserves the design philosophy of Rubin (2007) as it does not require the use of any outcome or treatment data when selecting a model for the instrument propensities which helps to avoid post-model-selection problems. A doubly robust extension is proposed as well. The balancing estimator is asymptotically normal and reaches the semiparametric efficiency bound if the number of balancing constraints grows.

ARTICLE HISTORY

Received July 2020
Accepted June 2021

KEYWORDS

Instrumental variable; Inverse probability weighting; Treatment effect

Accepted June 2021
appropriately with the sample size. Monte Carlo simulations suggest that the theoretical advantages over conventional methods translate well to finite samples. The method is applied to a re-evaluation of the causal effect of 401(k) participation on total financial assets.

The article is structured as follows: Section 2 discusses the related literature. Section 3 introduces the identification assumptions and the basic arguments behind balancing. Section 4 presents the estimation strategies. Section 5 provides the statistical properties of the balancing estimator. Section 6 discusses the choice of balancing functions in applications. Section 7 contains the Monte Carlo simulations. Section 8 provides the application and Section 9 concludes. All proofs and derivations are collected in the appendix.

2. Related Literature

Identification and non/semiparametric estimation of the LATE using a binary conditionally independent instrument has been first considered by Abadie (2003) and Frölich (2007). Abadie (2003) relied on inverse probability weights to identify conditional expectation functions for compliers. His method requires a model for complier outcomes and is thus more prone to misspecification. In practice, the method often yields estimates close to conventional two-stage least squares if linear models are used (Angrist and Pischke 2009). Frölich (2007) proposed nonparametric imputation/matching estimators and shows that the semiparametric efficiency is not affected by knowledge of the instrument propensity scores similar to Hahn (1998) in the context of selection on observables. Frölich (2007) also suggested the use of an IPW estimator but does not provide any theory for estimation. Donald, Hsu, and Lieli (2014b) proposed an IPW estimator for the LATE that relies on nonparametric series estimation for the instrument propensity scores similar to Hirano, Imbens, and Ridder (2003) and provided conditions for semiparametric efficiency, see also Donald, Hsu, and Lieli (2014a) for local polynomial regression. The IPW estimators in Frölich (2007), Donald, Hsu, and Lieli (2014a), and Donald, Hsu, and Lieli (2014b) are all of the "IPW1"-type, that is, they do not impose normalization of the inverse probability weights. Our approach is closest to Donald, Hsu, and Lieli (2014b) but we use balancing conditions instead of a series logistic estimator. However, the balancing estimator is automatically weight-normalized, favors moderate instrument propensities, and imposes exact mean balance in finite samples. Thus, despite asymptotic equivalence, there are major differences in bias and overall point estimation risk in finite samples.

Improving IPW estimation through imposing exact or approximate balancing constraints has been considered in the literature on estimation of treatment effect under selection on observables. Graham, de Xavier Pinto, and Egel (2012) considered tilted moment conditions for estimation of conventional propensity scores. Hainmueller (2012) proposes direct optimization of a distance criterion depending on inverse probability weights. Zhao and Percival (2017) provide conditions under which the method by Hainmueller (2012) is doubly robust for the average treatment effect on the treated. Imai and Ratkovic (2014) consider exact balancing of covariates in a parametric GMM and empirical likelihood framework. Zubizarreta (2015) proposes to minimize a quadratic problem in terms the inverse probability weights subject to approximate mean balancing constraints. Athey, Imbens, and Wager (2018) employ approximate balancing weights for bias-correction of high-dimensional linear models, see also Ning, Sida, and Imai (2020) for a doubly robust balancing method in high dimensions. Zhao (2019) develops a unifying framework for empirical balancing approaches and demonstrates how to tailor loss functions to produce weights that correspond to a treatment effect of interest. Many of these contributions suggest that empirical balancing can substantially outperform conventional weighting estimators by reducing differences between the weighted empirical distributions of treatment and control units. Our balancing approach in its most basic form is a repeated application of the exact balancing method by Imai and Ratkovic (2014) and Zhao (2019) using instrument instead of treatment indicators. However, it has different bias and weight normalization properties that do not apply in the context of selection on observables. Moreover for the LATE there is an extended causal mechanism of three different stages: instrument assignment, treatment selection, and outcome generation. We demonstrate that using information from the latter two steps can achieve approximately unbiased estimates and even double robustness if appropriate balancing conditions invoked along the lines of Fan et al. (2020).

Applications that rely on conditionally independent instruments are manifold, see, for example, Angrist (1990), Poterba, Venti, and Wise (1995), and Abadie (2003). A recent empirical article by Knaus, Lechner, and Reimers (2020) contained an idea similar to the one proposed in this article. They combine two steps of inverse probability tilting by Graham, de Xavier Pinto, and Egel (2012) to obtain a balanced LATE. However they do not provide any theory for estimation or statistical inference.

3. Identification and Balancing

We consider standard identification conditions for treatment effects under unobservable heterogeneity and a binary conditionally independent instrument often referred to as assignment. Assume that we observe independent data \((Y_i, D_i, Z_i, X_i')\) for units \(i = 1, \ldots, n\). \(X_i\) is a vector of (causally) predetermined random variables supported on \(\mathbb{R}^{\text{dim} X_i}\). \(Z_i\) is a binary instrument, \(D_i\) a binary treatment, and \(Y_i\) a real-valued outcome variable. If there are no interferences across units (i.e., SUTVA holds as in Angrist, Imbens, and Rubin 1996), then there are four potential outcome states \(Y_i(d, z)\) for \(d, z \in \{0, 1\}\) but only \(Y_i = Y_i(d_i, Z_i)\) is observed. For each instrument level \(z \in \{0, 1\}\), there is a potential treatment status \(D_i(z)\) which yields observed treatment \(D_i = Z_i D_i(1) + (1 - Z_i) D_i(0)\). The following identification assumptions along the lines of Abadie (2003) and Frölich (2007) are imposed, see also Donald, Hsu, and Lieli (2014a,b):

(A.1) (Conditional Independence) \(Y_i(d, z); \forall d, z, D_i(1), D_i(0) \perp \perp Z_i|X_i\).

(A.2) (Exclusion) \(P(Y_i(d, 1) = Y_i(d, 0)) = 1\) for \(d = 0, 1\).

(A.3) (Monotonicity) \(P(D_i(1) - D_i(0) \geq 0) = 1\).

(A.4) (First Stage) \(E[D_i(1) - D_i(0)] \neq 0\).
A.1–A.4 together are a weaker version of the conditions for identification of the LATE in the seminal article by Imbens and Angrist (1994) for completely independent instruments. Assumption A.1 is the fundamental identification assumption. It implies that after controlling for a sufficient set of observed confounding variables, any residual variation in potential outcomes and potential treatment statuses is independent of the instrument. Thus, conditional on observed confounders, the instrument can be thought of as being allocated like in a completely randomized experiment. Assumption A.2 rules out any direct effects of the instrument on potential outcomes other than through the indirect treatment channel. Thus, there cannot be any unobserved confounders affected by the instrument and no feedback from potential outcomes to the instrument. Under this exclusion, the observed outcome is given by $Y_i = D_i Y_i(1, Z_i) + (1 - D_i) Y_i(0, Z_i) = D_i Y_i(1) + (1 - D_i) Y_i(0)$. Assumption A.3 imposes a monotone effect of the instrument on the treatment choice, that is, receiving instrument $Z_i = 1$ makes any unit at least as likely to select itself or to be selected into treatment compared to $Z_i = 0$. This is sometimes referred to as the “no defiers” assumption (Angrist, Imbens, and Rubin 1996). Assumption A.4 assures that there is overall variation in potential treatment statuses as a result of variation in the instrument. Together with monotonicity, this requires the data to have a nonzero share of compliers, that is, units for which $D_i(1) > D_i(0)$. Assumption A.5 requires that potentially each unit could have been exposed to a different instrument level. Point identification only requires instrument overlap ($\delta = 0$). For regular behavior of point estimators, however, strong instrument overlap is often imposed, see, for example, Heiler and Kazak (2021). There are no further restrictions on the functional relationship between a unit’s potential outcomes, instrument, or covariates. Thus, the framework allows for almost any type of observable and unobservable heterogeneity in potential outcomes and causal effects $Y_i(1) - Y_i(0)$. Frölich (2007) showed that under similar assumptions as A.1–A.5, the local average treatment effect $tLATE = E[Y_i(1) - Y_i(0)|D_i(1) > D_i(0)]$ is nonparametrically identified. He only assumes conditional mean independence for identification but for root-n consistent estimation and asymptotic variance estimation higher moment independencies are necessary as well. $tLATE$ is the average treatment effect for the subpopulation of compliers, that is, the expected causal effect for a unit randomly drawn from the population of units that alter their potential treatment choice in accordance with the instrument. Without further assumptions, identification does not extend to causal effects such as the average treatment effect (ATE) or the treatment effect on the treated (TT). Under effect homogeneity, the LATE is equal to the ATE and the TT. In the case of one-sided noncompliance $P(D_i(0) = 1) = 0$, LATE equals TT (Bloom 1984). One-sided noncompliance often occurs in randomized field experiments with imperfect compliance if the treatment can only be provided by the experimenter.

The identification results in Abadie (2003) and Frölich (2007) allow for the construction of matching, model-based imputation, and inverse probability weighting estimators for the LATE. In this article, we focus on the latter as they can produce efficient estimates and only require estimation of the instrument propensity scores. Both treatment selection and in particular potential outcome mechanisms can be more difficult to model as they are often the results of complicated mechanisms such as markets, search and matching processes, or social and biological structures and interactions. Focusing on the instrument or assignment phase avoids misspecification and post-model-selection problems with statistical inference as outcome and treatment data do not have to be used for obtaining instrument proprieties. This parallels the argument for focusing on the design phase of experimental and observational studies (Rubin 2007).

In what follows, we outline the balancing principle behind inverse probability weighting and show how this allows us to extract population constraints that can be used for estimation of the instrument propensity scores. For exploiting identification via inverse probability weighting note that

$$t_{LATE} = E[Y_i(1) - Y_i(0)|D_i(1) > D_i(0)] = \Delta / \Gamma \quad (3.1)$$

with

$$\Delta = E \left[ \frac{Z_i Y_i}{\pi(x_i)} \right] - E \left[ \frac{(1 - Z_i) Y_i}{1 - \pi(x_i)} \right],$$

$$\Gamma = E \left[ \frac{Z_i D_i}{\pi(x_i)} \right] - E \left[ \frac{(1 - Z_i) D_i}{1 - \pi(x_i)} \right]. \quad (3.2)$$

Thus, both numerator and denominator of the LATE can be written as a difference of two expectations of inverse probability weighted quantities that can be identified from the joint distribution of $(Y_i, D_i, Z_i, X_i)$. The fundamental mechanism behind identification via inverse probability weighting is the balancing property, that is, the inverse probability weights will balance the distribution of any function of covariates across instrument levels. Let $f_2(x)$ denote the density of the observed confounders and $f_{Z_i=1}(x)$ the density of the observed confounders conditional on $Z_i = 1$. By Bayes’ Law it follows that for all $x \in X'$

$$f_{Z_i=1}(x) \pi(x) = f_{Z_i=0}(x) \pi(x) \quad (3.3)$$

Equation (3.3) implies a mean balancing property for any function of the covariates: Let $g : X' \rightarrow \mathbb{R}$ be a function with $E[|g(X_i)|] < \infty$. The mean of $g(\cdot)$ is balanced across inverse probability weighted instrument groups and corresponds to the unweighted population mean, that is,

$$E \left[ \frac{g(X_i) Z_i}{\pi(x_i)} \right] = E \left[ \frac{g(X_i)(1 - Z_i)}{1 - \pi(x_i)} \right] = E[g(X_i)]. \quad (3.4)$$

Thus, despite allowing for dependent unobservables driving potential outcome and treatment decision, balancing on observed confounders is enough to achieve causal identification for the compliers. This is due to the fact that conditional on observables the instrument is as good as randomly allocated. Thus, any imbalances in unobservables that would introduce a bias when comparing means between units from different treatment levels for always-takers, never-takers, and compliers combined do not matter for the (unidentified) population of compliers as they choose treatment in accordance with the instrument, that is, $D_i(z) = z$. The remaining difference due

(A.5) (Strong Instrument Overlap) Let $\pi(x) = P(Z_i = 1|X_i = x)$. There exists a $\delta > 0$ such that $\delta < \pi(x) < 1 - \delta$ for all $x \in X'$. It implies that after controlling for a sufficient set of observed confounding variables, any residual variation in potential outcomes and potential treatment statuses is independent of the instrument.
to varying treatment selection is then accounted for by the denominator in (3.1) that corrects for the share of compliers. Thus, balancing for compliers boils down to balancing two instrument assignment groups twice for different outcomes.

While population balance (3.4) is present when true instrument propensities are used, note that when using true scores or choosing a model for \( \pi(X_i) \) in finite samples, the ultimate goal is to impose balance and not to necessarily choose a model that best fits \( \pi(X_i) \) or \( Z_i \) in a given sample according to a standard loss function such as the log-likelihood. From a population perspective or asymptotically these goals are usually aligned. In finite samples, however, balance is not guaranteed and hence any differences between the weighted distributions can still substantially compromise estimation and causal inference. Thus to reduce bias choosing an estimator that exploits (3.4) directly should be beneficial.

4. Estimation
4.1. The Balancing Estimator

This section introduces the balancing estimator for the LATE. First, one has to choose a link function for the instrument propensities similar to likelihood based models. Second, a proper scoring rule tailored to the chosen link that guarantees finite sample balance is constructed. Its corresponding probability estimates are then used in an IPW estimator for the LATE. Let \( L : \mathbb{R} \to [0,1] \) denote a strictly monotone, continuously differentiable link function with derivative \( L' \). For now, instrument propensity scores are parameterized by a single-index \( \pi(X_i, \theta) = L(\phi(X_i)'\theta) \) where \( \phi : \mathcal{X} \to \mathbb{R}^r \) with \( r < n \). For an extension to non-parametric smooth function classes consider Section 5.3. The propensities that balance covariates \( \phi(X_i) = (\phi_1(X_i), \ldots, \phi_r(X_i))' \) are then obtained by maximizing the (negative) tailored loss function

\[
\hat{\theta} = \arg \max_{\theta} -\frac{1}{n} \sum_{i=1}^{n} \text{S}(Z_i, X_i, \theta)
\]

\[
= \arg \max_{\theta} -\frac{1}{n} \sum_{i=1}^{n} Z_i S_1(\phi(X_i)'\theta) + (1 - Z_i) S_0(\phi(X_i)'\theta),
\]

(4.1)

where

\[
S_1(x) = \int \frac{1}{L(x) L'(x)} dL(x),
\]

\[
S_0(x) = -\int \frac{1}{(1 - L(x)) L'(x)} dL(x).
\]

(4.2)

\( S(\cdot) \) matches the definition of a proper scoring rule for the binomial variable \( Z_i \) (Savage 1971). Thus, the solution to Equation (4.1) yields Fisher consistent estimates for the instrument propensity scores if the link is correctly specified. Note that the loss function explicitly depends on the link function to assure balancing and global concavity. In comparison to a log-likelihood function it has different sensitivity to extreme propensity scores as it is tilted around the maximizers. Consider, for example, the logistic link where

\[
S_1(\phi(X_i)'\theta) = \phi(X_i)'\theta - 1/L(\phi(X_i)'\theta)
\]

\[
S_0(\phi(X_i)'\theta) = -\phi(X_i)'\theta - 1/(1 - L(\phi(X_i)'\theta)).
\]

(4.3)

One can see that, for \( z = 1 \), the log-odds ratio is maximized while also penalizing small propensity scores and conversely for \( z = 0 \). Setting the first-order condition of (4.1) for any link function to zero yields

\[
\frac{1}{n} \sum_{i=1}^{n} \left[ \frac{Z_i}{L(\phi(X_i)'\theta)} - \frac{1 - Z_i}{1 - L(\phi(X_i)'\theta)} \right] \phi(X_i) = 0. \tag{4.4}
\]

This explicitly exploits the sample equivalents of Equation (3.4) and yields a transparent connection between the choice of balancing functions and the choice of regressors in the instrument propensity score model. It corresponds to the estimator by Imai and Ratkovic (2014) in the context of selection on observables with instrument replaced by treatment. It is central to note that the validity of the tailored loss approach for estimating the LATE does not hinge on the correct specification of the link function if the right balancing functions are chosen, see Section 5.1. Let \( \hat{\pi}(X_i) = L(\phi(X_i)'\hat{\theta}) \). The balancing estimator for the LATE is then given by

\[
\hat{\tau}_{\text{LATE}} = \hat{\Delta}/\hat{\Gamma}
\]

(4.5)

with

\[
\hat{\Delta} = \frac{1}{n} \sum_{i=1}^{n} Z_i Y_i \frac{1}{\hat{\pi}(X_i)} - \frac{1}{n} \sum_{i=1}^{n} (1 - Z_i) Y_i \frac{1}{\hat{\pi}(X_i)},
\]

\[
\hat{\Gamma} = \frac{1}{n} \sum_{i=1}^{n} Z_i D_i \frac{1}{\hat{\pi}(X_i)} - \frac{1}{n} \sum_{i=1}^{n} (1 - Z_i) D_i \frac{1}{1 - \hat{\pi}(X_i)}.
\]

(4.6)

A unique feature of Equation (4.4) is that if \( \phi(X_i) \) contains an intercept, then estimator (4.5) is automatically weight-normalized, that is, numerically identical to its "IPW2" version that uses weights \( Z_i/\hat{\pi}(X_i)/[n^{-1} \sum_{i=1}^{n} Z_i/\hat{\pi}(X_i)] \) and \( (1 - Z_i)/(1 - \hat{\pi}(X_i))/[n^{-1} \sum_{i=1}^{n} (1 - Z_i)/(1 - \hat{\pi}(X_i))] \). This follows from the ratio form of \( \hat{\tau}_{\text{LATE}} \) together with Equation (4.4) and \( \phi(X_i) = c \) for any \( c \neq 0 \). It is not the case that balanced inverse probability weights themselves are normalized to unity. Under selection on observables, there is clear evidence that weight-normalized versions of IPW estimators generally outperform their unweighted counterparts due to reduction in variance (Busso, DiNardo, and McCrary 2014; Pohlmeier, Seiberlich, and Uysal 2016).

4.2. Doubly Robust Extended Balancing

A possible extension is to augment balancing conditions (4.4) with a second set of conditions that also balances the reweighted covariates in the treatment group with the unweighted covariates in the control group. This can reduce finite sample bias under weaker assumptions compared to standard balancing. To do so, parameterize the instrument scores as \( L(\phi(X_i)'\theta_1 + \phi(X_i)'\theta_2) \) where \( \phi : \mathcal{X} \to \mathbb{R}^2 \) such that \( r + r_2 < n \). The extended balancing instrument propensity scores are then given by \( \hat{\pi}(X_i) = L(\phi(X_i)'\theta_1 + \phi(X_i)'\theta_2) \) where \( \theta_1 \) and \( \theta_2 \) are the...
solution to the following sample moment equations:

\[
\frac{1}{n} \sum_{i=1}^{n} \left[ \frac{Z_i}{L(\phi(X_i)^{\theta_1} + \phi(X_i)^{\theta_2})} - 1 \right] \phi(X_i) = 0, \\
\frac{1}{n} \sum_{i=1}^{n} \left[ \frac{Z_i}{L(\phi(X_i)^{\theta_1} + \phi(X_i)^{\theta_2})} - 1 \right] \hat{\phi}(X_i) = 0. \tag{4.7}
\]

These are \(r + r_2\) equations for the same number of parameters. A recent working article by Fan et al. (2020) on ATE estimation demonstrates that combining these two balancing conditions can be exploited to achieve doubly robust versions of covariate balancing estimators. Their method is motivated by removing the dependence of optimal balancing functions from the explicit propensity score model to contain elements that capture variation in the conditional mean of the control outcome score model to contain elements that capture variation in the conditional mean of the causal effects of compliers and always-takers. For the extended balancing estimator, finite sample robustness is achieved by incorporating additional effect and selection information into the instrument propensity score model. Note that this also implies a full double robustness property in large samples, that is, either the propensity score has to be correctly specified or the span conditions (ii) have to be matched in large samples for asymptotically unbiased estimates. This is similar to augmented IPW estimation, where the potential outcomes models incorporate this information (Robins and Rotnitzky 1995). Note that for IPW the finite sample bias property in Proposition 5.1 is unique to balanced instrument propensity scores generally does not apply to other estimation approaches such as maximum likelihood or even true instrument propensity scores. As the ratio of the expectations is only an approximation to the expectation of the ratio in finite samples, the actual finite sample bias has to be further investigated. We return to this point in Section 6.

### 5.2. Duality and Variance Reduction

In this section, we show that in finite samples the balancing approach favors moderate instrument propensity scores in the sense of being close to one half. As instrument propensity scores are inversely related to the (conditional) variance of IPW estimators, there are potential point estimation gains compared to the expectation of the ratio in finite samples, the actual finite sample bias has to be further investigated. We return to this point in Section 6.

The robustness property for the regular balancing estimator (i) is best understood in comparison with the two-stage least squares (2SLS) estimator. If a first stage for the endogenous treatment variable is fully saturated, then the model behind 2SLS identifies a weighted version of conditional LATEs with weights being proportional to the conditional variances of the first stages \(V[E[D_i(Z_i)|X_i] \text{, see Angrist and Imbens (1995). Under homogeneous complier effects this corresponds to the LATE. The parameters from the reduced forms for both outcome and endogenous variable can then be estimated without bias. Thus, 2SLS is a ratio of two unbiased estimators with the ratio of the expectations being equal to the LATE. Proposition 5.1 states that the balanced IPW estimator has an equivalent property under comparable assumptions. However, instead of having a flexible model for the treatment choice \(E[D_i(X_i,Z_i)]\), the flexibility is incorporated through the choice of \(\phi(X_i)\) in the instrument scores.

For the extended balancing estimator finite sample robustness is achieved by incorporating additional effect and selection information into the instrument propensity score model. Note that this also implies a full double robustness property in large samples, that is, either the propensity score has to be correctly specified or the span conditions (ii) have to be matched in large samples for asymptotically unbiased estimates. This is similar to augmented IPW estimation, where the potential outcomes models incorporate this information (Robins and Rotnitzky 1995). Note that for IPW the finite sample bias property in Proposition 5.1 is unique to balanced instrument propensity scores generally does not apply to other estimation approaches such as maximum likelihood or even true instrument propensity scores. As the ratio of the expectations is only an approximation to the expectation of the ratio in finite samples, the actual finite sample bias has to be further investigated. We return to this point in Section 7.

### 5. Statistical Properties of the Balancing Estimator

#### 5.1. Approximate Finite Sample Bias

The balancing estimators yield asymptotically unbiased estimates of the LATE if the link function is correctly specified. This section demonstrates that, with appropriate choice of the balancing estimators, there are potential point estimation gains compared to augmented IPW estimation, where the potential outcomes models incorporate this information (Robins and Rotnitzky 1995). Note that for IPW the finite sample bias property in Proposition 5.1 is unique to balanced instrument propensity scores generally does not apply to other estimation approaches such as maximum likelihood or even true instrument propensity scores. As the ratio of the expectations is only an approximation to the expectation of the ratio in finite samples, the actual finite sample bias has to be further investigated. We return to this point in Section 7.

The first span condition requires the instrument propensity score model to contain elements that capture variation in the conditional mean of the control outcome \(E[Y_i(0)|X_i]\). Moreover, balancing components have to be flexible enough to cover the relevant building blocks for the conditional mean of a potential treatment level and the conditional causal effect for a combination of different units. These conditions can be reformulated by using the monotonicity assumption. For \(z = 0\), it reduces to the assumption that \(E[D_i(0)|X_i]\) and the product between \(E[D_i(z)|X_i] \text{ and } E[D_i(0)|X_i]\) and the conditional causal effect for the always-takers \(\tau_{AT}(X_i)\) and \(\tau_{LATE}(X_i)\) to be in the linear span as \(E[D_i(0)(Y_1(1) - Y_0(1))|X_i] = \tau_{LATE}(X_i)E[D_i(0)|X_i]\). For \(z = 1\), it requires a linear combination of probability weighted versions of the causal effects of compliers and always-takers to be captured since \(E[D_i(1)(Y_1(1) - Y_0(1))|X_i] = \tau_{LATE}(X_i)E[D_i(1) - D_i(0)|X_i] + \tau_{AT}(X_i)E[D_i(0)|X_i]\). Thus, low bias demands a special type of flexibility regarding the transformations of regressors \(\phi(X_i)\) chosen for balancing. Depending on the application at hand, this can be restrictive. The derivations for Proposition 5.1 reveal that it is actually not necessary that the instrument scores used in the denominator have the same degree of flexibility as the ones in the numerator. In particular, for denominator scores only a condition for \(E[D_i(z)|X_i] \text{ for } z \in \{0, 1\}\) is required. This can be exploited by choosing different instrument propensity score models for numerator and denominator. We return to this point in Section 6.

The robustness property for the regular balancing estimator (i) is best understood in comparison with the two-stage least squares (2SLS) estimator. If a first stage for the endogenous treatment variable is fully saturated, then the model behind 2SLS identifies a weighted version of conditional LATEs with weights being proportional to the conditional variances of the first stages \(V[E[D_i(Z_i)|X_i] \text{, see Angrist and Imbens (1995). Under homogeneous complier effects this corresponds to the LATE. The parameters from the reduced forms for both outcome and endogenous variable can then be estimated without bias. Thus, 2SLS is a ratio of two unbiased estimators with the ratio of the expectations being equal to the LATE. Proposition 5.1 states that the balanced IPW estimator has an equivalent property under comparable assumptions. However, instead of having a flexible model for the treatment choice \(E[D_i(X_i,Z_i)]\), the flexibility is incorporated through the choice of \(\phi(X_i)\) in the instrument scores.
to using likelihood scores. For moderate inverse probability weights, maximizing (4.1) penalizes deviations from the unconditional mean proportional to the conditional variance of the LATE estimator. Let the inverse probability weights be defined as

$$w_i = \frac{Z_i}{\pi(X_i)} + \frac{1 - Z_i}{1 - \pi(X_i)}$$  \hspace{1cm} (5.2)

and denote $W = (W_1, \ldots, W_n)$ with $W_i = (X_i', Z_i)$ for $i = 1, \ldots, n$. Conditional on $W$, the balanced instrument propensity scores are known. Thus, a second order Taylor expansion of the variance of the balanced LATE (4.5) conditional on $W$ yields

$$nV\left[\frac{\hat{\Delta}}{W}\right] \approx \frac{1}{n} \sum_{i=1}^{n} w_i^2 a(X_i, Z_i)$$  \hspace{1cm} (5.3)

with

$$a(X_i, Z_i) = \left( V[Y_i|W_i]E[\hat{\Delta}|W] \right)^2 - 2\text{cov}(Y_i, D_i|W_i)E[\hat{\Delta}|W]E[\hat{\Delta}|W] + E[D_i|W_i]E[\hat{\Delta}|W] \right)^2 E[\hat{\Delta}|W]^{-4}$$  \hspace{1cm} (5.4)

which is strictly greater than zero. Hence, the conditional variance is bounded from above by $\sup_x a(x, z) \frac{1}{n} \sum_{i=1}^{n} w_i^2$. It is directly proportional to the squared inverse probability weights under homoscedasticity. We assume that $\sum_{i=1}^{n} w_i^2 < \infty$. For $w_i > 0$ and $w_i \neq 1$, this implies that balancing covariates without perfect randomization are generally opposing goals. In principal, one could also use the characterization of the conditional variance in Equation (5.4) to choose minimizing weights within a class of balancing weights as proposed in the context of selection on observables by Li, Morgan, and Zaslavsky (2018). This strategy, however, changes the definition of the underlying identified causal parameter from the LATE to a ratio of two weighted reduced form estimates which put a higher weight on individual with instrument propensities close to one half. The resulting identified parameter does not lend itself to an intuitive causal interpretation with similar policy relevance compared to the conventional LATE.

### 5.3. Nonparametric Estimation and Large Sample Properties

In light of Proposition 5.1, it seems reasonable to choose a model for the instrument propensity score that eventually incorporates a large set of balancing constraints as the sample size increases. In this section, we show that using a nonparametric approach within the tailored loss framework can efficiently incorporate all required information for estimation of the LATE. It does so while still retaining the exact finite sample balancing property in (4.4). We rely on a series approach using power series similar to Hirano, Imbens, and Ridder (2003) and Donald, Hsu, and Lieli (2014b) who relied on a local maximum likelihood step for estimation of the scores. For $K > 0$, let $\phi^K(x) = (x^{\lambda(1)}, \ldots, x^{\lambda(K)})$ be a vector of power functions such that $|\lambda(k)| \leq |\lambda(k + 1)|$ for $k \in \mathbb{N}_0$ with $\lambda = (\lambda_1, \ldots, \lambda_r)'$, nonnegative and $|\lambda| \leq 1$ denoting the $\ell_1$-norm, that is, $|\lambda|_1 = \sum_{i=1}^{r} |\lambda_i|$. We assume that the series is orthogonalized with respect to a weight function such that $E[\phi^K(X_i)\phi^K(X_i)'] = I_K$. This is always possible since a logistic link function is used and thus we approximate the log-odds ratio by a linear single index $\phi^K(x)'\theta_K = \theta_K^0 A^{-1}_K A \phi^K(x)$. Therefore, one can always use basis $A_k \phi^K(x)$ for approximation instead, see also Appendix A in Hirano, Imbens, and Ridder (2003). The balanced instrument propensity scores are then given by $\hat{\pi}(x) = L(\phi^K(X_i)^)'$ with

$$\hat{\theta}_K = \arg \max_{\theta} \frac{1}{n} \sum_{i=1}^{n} S(Z_i, X_i, \theta),$$  \hspace{1cm} (5.7)

$$S(Z_i, X_i, \theta) = (2Z_i - 1)\phi^K(X_i)^' \theta - \frac{Z_i}{L(\phi^K(X_i)^')} - \frac{1 - Z_i}{1 - L(\phi^K(X_i)^')}.$$  \hspace{1cm} (5.8)
Let $C^d$ denote the space of $d$-times continuously differentiable functions. We impose the following regularity and smoothness assumptions:

(B.1) $X_i$ is an $r$-dimensional random vector compactly supported on $\mathcal{X}$ with absolutely continuous density $f(x)$ in $C^2$ and bounded away from zero.

(B.2) $E[Y_i|X_i, Z_i = z] = m_z(X_i)$ and $E[D_i|X_i, Z_i = z] = \mu_z(X_i)$ are in $C^4$.

(B.3) $\pi(X_i)$ are in $C^3$ with $q \geq 7r$.

(B.4) All second moments of potential outcome levels exist and are finite.

(B.5) $K = O(n^r)$ with $1/(4(q/r - 1)) < v < 1/9$.

The assumptions are standard in the literature, see Hirano, Imbens, and Ridder (2003) and Donald, Hsu, and Lieli (2014b) or Li, Racine, and Wooldridge (2009) and Donald, Hsu, and Lieli (2014a,b) for possible restrictions on the series terms and adaptations to discrete covariates and kernel methods. Assumption B.1 assures a uniform approximation of any continuous function of the covariates, in particular conditional means such as the instrument propensity score. Assumptions B.2 and B.3 are smoothness conditions on the observed outcome and treatment status conditional on covariates and instrument level and on the instrument propensity score. The higher the dimensionality of the covariates, the more smoothness conditions are required. B.4 is a regularity condition that is necessary to assure a finite asymptotic variance for the LATE estimator. Assumption B.5 controls the rate at which the order of basis functions is allowed to grow with increasing sample size depending on the degree of smoothness. We obtain the following theorem:

**Theorem 5.1 (Efficient Balancing).** Under Assumptions A.1–A.5, B.1–B.5, and instrument propensity scores estimated according to Equation (5.7), the inverse probability weighting estimator for the LATE

1. is asymptotically normal

$$\sqrt{n}(\hat{\tau}_{LATE} - \tau_{LATE}) \overset{d}{\rightarrow} \mathcal{N}(0, V)$$

2. and reaches the semiparametric efficiency bound

$$V = \frac{1}{n^2} \left( E \left[ (m_1(X_i) - m_0(X_i) - \tau_{LATE} \mu_1(X_i) + \tau_{LATE} \mu_0(X_i))^2 \right] 
+ \sum_{z=0,1} E \left[ \sigma_{Y_i,D_i}^2(X_i) - 2\tau_{LATE} \sigma_{Y_i,D_i}(X_i) + \tau_{LATE}^2 \sigma_{D_i}^2(X_i) \right] \right)$$

with $\sigma_{Y_i,D_i}^2(X_i) = V[Y_i|X_i, Z_i = z]$, $\sigma_{D_i}^2(X_i) = V[D_i|X_i, Z_i = z]$, and $\sigma_{Y_i,D_i}(X_i) = \text{cov}[Y_i, D_i|X_i, Z_i = z]$ for any $z \in \{0, 1\}$.

Theorem 5.1 shows that the inverse probability weighting estimator using a sufficiently flexible nonparametric model for the instrument propensity score that imposes empirical balancing constraints efficiently incorporates all information available for estimation of the LATE and is consistent. This is in line with the insights from the bias characterizations in Section 5.1. There, the instrument propensity scores yield first-order unbiased estimates if they incorporate the structure of the potential treatment effects and potential treatment levels sufficiently. If these conditional mean functions are continuous, then a series approximation will eventually contain all the components necessary to uniformly approximate them on compact sets. Theorem 5.1 also implies that in large samples there is no qualitative difference between tailored instrument propensity scores and standard nonparametric approaches. However, the balancing approach additionally guarantees finite sample balance. The efficiency bound for the LATE has originally been derived by Frölich (2007), see also Hong and Nekipelov (2010). The asymptotic variance can be consistently estimated using a series approach as in Hirano, Imbens, and Ridder (2003) and Donald, Hsu, and Lieli (2014b) by replacing the population quantities with sample estimates, see supplementary material. Recent work by Fan et al. (2020) on ATE estimation suggests that for extended balancing (4.7) smoothness Assumption B.3 could potentially be weakened to $q > 0.5r$ without sacrificing asymptotic normality and semiparametric efficiency. We leave an extension for the LATE along this line for future work.

### 6. Selecting Balancing Functions

The results in Section 5.3 make a strong case for eventual flexibility of the balancing variables. In large samples, however, the exact choice or order of inclusion of transformations of regressors is not of primary concern. Thus, the question remains, which empirical means should be prioritized for balancing in finite samples, that is, how to pick $\phi(X_i)$? From a model selection perspective, choosing informative balancing variables first should be beneficial. From an applied perspective, it is often reasonable to assume varying degrees of knowledge about the different steps of the underlying causal mechanism from (i) instrument assignment over (ii) treatment choice to (iii) outcome. Consider the example of evaluating the impact of a job search training program on future earnings for the unemployed offered by an agency: The decision to assign units could be based on a set of observable characteristics such as age, employment history, and qualifications available to the agent responsible for assignment. In (ii), the units choose whether to comply with the assignment. If enough information about their tradeoffs and restrictions are available, the choice problem and its constraints can be modeled guided by economic theory. The outcome process determining earnings, however, is likely generated as a consequence of a complicated searching and matching process on the labor market under additional constraints. Thus, relying only on data from the employment agency might not be enough to really inform a model for (iii).

In light of Proposition 5.1, balancing (a) $E[Y_i(0)|X_i], E[D_i(z)|Y_i(1) - Y_i(0)|X_i]$ and (b) $E[D_i(z)|X_i]$ would in principle be desirable for a $z \in [0, 1]$. (a) requires outcome data to generate, for example, model-based quantities. Hence, using them for empirical balancing operates on level (i) going against the design arguments outlined by Rubin (2007). (b) on the other hand is a better candidate as it only concerns treatment choice. For simplicity assume that $X_i$ is discrete. From Vytlacil (2002), it follows that Assumptions A.1–A.4 (conditional on $X_i$) imply the existence a nonparametric single index model that rationalizes the identical choices as the LATE framework conditional on each $x \in \mathcal{X}$ and vice versa. This motivates a fully nonparametric model for the first stage

$$D_i = 1(\mu(X_i, Z_i) > v_i) \quad (6.1)$$
with \( v_i \perp\!\!\!\perp Z_i | X_i, \mu(x, z) \) nondegenerate conditional on \( x \), and \( F_{v|x}(v) \) continuous. Without any further restrictions this corresponds to the first stage of a fully saturated instrumental variables approach. The estimates for \( E[D_i(z)|X_i] \) are obtained from the estimated \( F_{v|x}(\mu(x, z)) \). They can then be included into \( \phi(X) \) to balance the empirical counterparts of \( E[D_i(z)|X_i] \). For estimation of Equation (6.1) with many categories or multiple discrete regressors smoothing methods as in Ouyang, Li, and Racine (2009) or Heiler and Mareckova (2021) might be desirable. The proof of Proposition 5.1 reveals that this strategy is sufficient for the denominator to be unbiased. For the numerator, however, information about the outcome process would be required. If one is willing to impose a model for the potential outcomes, then from a bias perspective it would be sufficient to include them into the balancing constraints for the instrument propensity scores that enter the numerator only. Thus, one can operate with two different instrument propensity scores that differ by balancing the empirical counterparts of either \( E[D_i(z)|X_i] \) (denominator) or \( E[Y_i(0)|X_i] \) and \( E[D_i(z)|Y_i(1) - Y_i(0)|X_i] \) (numerator). Under constant treatment effects for some \( z \in \{0, 1\} \), that is, for always-takers and compliers or never-takers and compliers \( E[D_i(z)|Y_i(1) - Y_i(0)|X_i] = E[D_i(z)|X_i] \). Thus, the second component in (i) is balanced by simply including \( E[D_i(z)|X_i] \) into the instrument propensity score model. Therefore, incorporating \( E[D_i(z)|X_i] \) for balancing seems like a reasonable middle-ground between a fully agnostic approach and imposing a lot of structure on the outcome process.

If a parametric model for \( \mu(X_i, Z_i) \) is used, then there are two additional aspects to consider. First, statistical inference has to be adjusted to the presence of generated regressors. Second, balancing choices \( E[D_i(1)|X_i] \) and \( E[D_i(0)|X_i] \) are potentially highly correlated if variation of the instrument only has a mild effect on treatment selection and the expected share of compliers is small. In general, one of these components is enough for approximate unbiasedness.

### 7. Monte Carlo Study

#### 7.1. Design

In this section, we compare the finite sample performance of the balancing estimator and some of the proposed extensions to standard estimation approaches from the literature. Sections 5.1 and 5.2 provide some reasoning why reductions in points estimation risk from balancing could be due to both bias and variance. The simulations aim at disentangling these driving forces. We also evaluate the impact of using model-based information from the treatment selection step for balancing as proposed in Section 6. Table 1 contains the Monte Carlo design in the spirit of a generalized Roy model, see also Heckman and Vytlacil (2005) for a simulation of a similar design using a continuous instrument.

| \( X_i \sim \text{Uniform}(0,1) \) | \( \pi(X_i) = 1/(1 + \exp(-\mu_2(X_i)\theta_2)) \) |
|-----------------------------|----------------------------------|
| \( Z_i \equiv I(u_i < \pi(X_i)) \) | \( u_i \sim \text{Uniform}(0,1) \) |
| \( D_i(x) = I(\mu_d(X_i, Z_i) > v_i) \) | \( \varepsilon(1) \sim \mathcal{N}(0, 1) \) |
| \( Y_i(1) = \mu_{y_1}(X_i) + \varepsilon(1) \) | \( \varepsilon(0) \sim \mathcal{N}(0, 0) \) |
| \( Y_i(0) = \varepsilon(0) \) | \( \mu_d(x, z) \) with \( \rho = 0.5, \theta_0 = \ln((1 - \delta)/\delta) \) and \( \mu_d(\cdot), \mu_{y_1}(\cdot) \), and \( \mu_2(\cdot) \) see Table 2. \( \theta_0 \) controls the degree of overlap based by imposing bounds \((\delta, 1 - \delta)\) on the instrument scores. \( \rho \neq 0 \) allows for selection on unobservables. We assume a nonzero correlation only between unobservables driving potential treatment outcome and treatment selection to simplify the analysis. This is without loss of generality as in the generalized Roy model under normality the functional form of the treatment effect is not affected by this choice up to a scaling factor. The model implies an additive separable contributions of observables and unobservables to the LATE, that is,

\[
\tau_{\text{LATE}} = \frac{E[\mu_{y_1}(X_i)F(\mu_d(X_i, 1)) - F(\mu_d(X_i, 0))]}{E[\mu_d(X_i, 1) - F(\mu_d(X_i, 0))]} + \rho \frac{E[\mu_d(X_i, 0) - f(\mu_d(X_i, 1))]}{E[\mu_d(X_i, 1) - F(\mu_d(X_i, 0))]} \tag{7.1}
\]

with \( F(\cdot) \) and \( f(\cdot) \) being the cumulative distribution function and the density function of the univariate standard normal distribution, respectively. Table 2 contains the different design specifications:

The linear index for treatment choice with the given parameters imposes monotonicity and share of compliers \( E[D_i(1) - D_i(0)] = 0.5 \) for all designs. In Designs A and B, heterogeneity in causal effects is achieved through the correlation \( \rho \) between unobservables \( \varepsilon_1(1) \) and \( v_i \) only. For Designs C and D, heterogeneity additionally stems from a direct contribution of observables to the potential outcome one. Design A represents the special case of a fully independent instrument as the observed confounder does not affect treatment choice. It also corresponds to a controlled randomized experiment with close to one-sided noncompliance as \( P(D_i = 1|Z_i = 1) > 0.9999 \). The homogeneous mean function for the outcome assures that the unconditional LATE in Equation (7.1) is composed of two equally sized components, that is, the overall contributions of the potential outcome mean and the “control function” part are identical. Design B is a conditionally independent design with homogeneous potential outcome means. The parameters are chosen to be favorable toward linear IV with exogenous covariates \( X_i \) and instrument \( Z_i \). Design C is a conditionally independent design with a nonlinear potential outcome mean. In this design, the semiparametric approaches that do not require specification of an outcome model still yield asymptotically unbiased results. The precise functional form of the potential outcome mean is not crucial. In general, other nonlinear mean functions with sufficient heterogeneity will produce qualitatively similar results. Design D is similar to C but also requires balancing of a quadratic term.

Table 3 contains all the considered estimation approaches. For additional simulations using alternative balancing functions and misspecification, see supplementary material. IV denotes
Table 3. Monte Carlo study: estimation methods.

| Name       | Description                                                                 |
|------------|-----------------------------------------------------------------------------|
| IV         | Instrumental variables estimator using binary instrument \( Z_i \) and additional control \( X_i \). |
| IPW1       | IPW estimator for the LATE using correctly specified maximum likelihood instrument propensity scores. |
| IPW2       | IPW1 with weight normalization.                                              |
| BAL1       | Balanced IPW with \( \phi(X_i) = (1 \mathbb{1}_i)^{\gamma} \).              |
| BAL2       | Balanced IPW with \( \phi(X_i) = (1 \mathbb{1}_i \mathbb{E}(D_i(0)|X_i)^{\gamma}) \) with \( \mathbb{E}(D_i(0)|X_i) \) obtained from a correctly specified probit. |
| BAL3       | Balanced IPW with \( \phi(X_i) = (1 X_i \mathbb{E}(D_i(0)|X_i))^{\gamma} \) with \( \mathbb{E}(D_i(0)|X_i) \) obtained from a correctly specified probit. |
| BAL4       | Balanced IPW with \( \phi(X_i) = (1 X_i \mathbb{E}(D_i(0)|X_i) \mathbb{E}(D_i(0)Y(0) - Y_i(0)|X_i))^{\gamma} \) with \( \mathbb{E}(D_i(0)|X_i) \) obtained from a correctly specified probit and \( \mathbb{E}(D_i(0)Y(0) - Y_i(0)|X_i) \) from a correctly specified Heckman selection model. |
| eBAL       | Extendend Balanced IPW with \( \phi(X_i) \) as BAL4 and \( \phi(X_i) = (1 \mathbb{E}(D_i(1) - D_i(0)|X_i) \mathbb{E}(D_i(1)Y(1) - Y_i(0)|X_i))^{\gamma} \). |
| B2m        | Balanced IPW with \( \phi(X_i) = (1 \mathbb{E}(D_i(0)|X_i))^{\gamma} \) with \( \mathbb{E}(D_i(0)|X_i) \) obtained from a misspecified logit model. |
| B3m        | Balanced IPW with \( \phi(X_i) = (1 X_i \mathbb{E}(D_i(0)|X_i))^{\gamma} \) with \( \mathbb{E}(D_i(0)|X_i) \) obtained from a misspecified logit model. |

Table 4. Design A: relative mean squared errors and absolute biases.

| Name | Description |
|------|-------------|
| IV   |                         |
| IPW1 |                         |
| IPW2 |                         |
| BAL1 | Balanced IPW with \( \phi(X_i) = (1 X_i)^{\gamma} \).              |
| BAL2 | Balanced IPW with \( \phi(X_i) = (1 \mathbb{1}_i \mathbb{E}(D_i(0)|X_i)^{\gamma}) \) with \( \mathbb{E}(D_i(0)|X_i) \) obtained from a correctly specified probit. |
| BAL3 | Balanced IPW with \( \phi(X_i) = (1 X_i \mathbb{E}(D_i(0)|X_i))^{\gamma} \) with \( \mathbb{E}(D_i(0)|X_i) \) obtained from a correctly specified probit. |
| BAL4 | Balanced IPW with \( \phi(X_i) = (1 X_i \mathbb{E}(D_i(0)|X_i) \mathbb{E}(D_i(0)Y(0) - Y_i(0)|X_i))^{\gamma} \) with \( \mathbb{E}(D_i(0)|X_i) \) obtained from a correctly specified probit and \( \mathbb{E}(D_i(0)Y(0) - Y_i(0)|X_i) \) from a correctly specified Heckman selection model. |
| eBAL | Extendend Balanced IPW with \( \phi(X_i) \) as BAL4 and \( \phi(X_i) = (1 \mathbb{E}(D_i(1) - D_i(0)|X_i) \mathbb{E}(D_i(1)Y(1) - Y_i(0)|X_i))^{\gamma} \). |
| B2m  | Balanced IPW with \( \phi(X_i) = (1 \mathbb{E}(D_i(0)|X_i))^{\gamma} \) with \( \mathbb{E}(D_i(0)|X_i) \) obtained from a misspecified logit model. |
| B3m  | Balanced IPW with \( \phi(X_i) = (1 X_i \mathbb{E}(D_i(0)|X_i))^{\gamma} \) with \( \mathbb{E}(D_i(0)|X_i) \) obtained from a misspecified logit model. |

Table 3–7 contain the mean squared errors and absolute biases for Designs A, B, C, and D for all estimation approaches from Table 3. Mean squared errors are normalized by the MSE of the IV estimator. Results are obtained from an experiment using 10,000 Monte Carlo replications with overlap parameters \( \delta = 0.01, 0.02, 0.05 \) and sample sizes \( n = 500, 1000 \) yielding an expected number of 250 and 500 complier units, respectively.

Overall, the results suggest that the unnormalized MLE approach is outclassed by all other methods due to its large finite sample variance. In general, most balancing approaches outperform the conventional MLE-based IPW estimators, IPW1 and IPW2, by a substantial margin depending on the design. The differences are most pronounced for small sample sizes and small \( \delta \). Exploiting information about the selection step is generally helpful to reduce points estimation risk. In particular, the balancing estimator using only the selection probabilities BAL2 consistently outperforms all other methods. The balancing estimator using both regressor and selection probabilities BAL3 also performs well but closer to the balancing estimator

the standard linear instrumental variables estimator using \( X_i \) as exogenous variable in the first-stage and the outcome model. IPW1 and IPW2 are the Horvitz–Thompson estimators for the LATE using likelihood instrument propensities without (IPW1) and with (IPW2) normalization of the inverse probability weights. BAL1 is the balancing estimator that uses the same regressors as the maximum likelihood approaches. BAL2 and BAL3 partially exploit the conditions for approximate unbiasedness in Proposition 5.1 by using estimates for \( \mathbb{E}(D_i(0)|X_i) \) as additional (or only) variable for balancing. BAL4 is similar to BAL3 but contains estimates for all balancing functions in Proposition 5.1. eBAL is the extended balancing estimator with all balancing functions from Proposition 5.1. B2m and B3m are similar to BAL2 and BAL3 but use a misspecified logit model instead of the true probit model for the generated regressor. Note that incorrect specifications for the treatment selection as additional balancing constraints do not affect the asymptotic validity of balancing as even a misspecified quantity as a function of \( X_i \) still leads to a valid balancing constraint in the sense of Equation (3.4), however not necessarily to one that minimizes bias.

7.2. Results
Table 5. Design B: relative mean squared errors and absolute biases.

| $\delta$ | $n$ | MSE | IPW1 | IPW2 | BAL1 | BAL2 | BAL3 | BAL4 | eBAL | B2m | B3m |
|----------|-----|-----|------|------|------|------|------|------|------|-----|-----|
| 0.01     | 500 | 1.00 | 0.0625 | 0.0533 | 500.3 | 3.7221 | 3.7221 | 2.7977 | 6.8297 | 140.19 | 2.7235 | 2.3216 | 271e3 |
|          | 1000| 1.00 | 0.0230 | 0.0228 | 2.6449 | 2.6230 | 2.4177 | 2.4177 | 2.4177 | 2.4177 | 2.4177 | 2.4177 | 2.4177 |
| 0.02     | 500 | 1.00 | 0.0363 | 0.0368 | 8.7654 | 1.8417 | 1.8417 | 2.1975 | 2.4710 | 2.8217 | 2.7235 | 2.3216 | 22067 |
|          | 1000| 1.00 | 0.0122 | 0.0122 | 2.1356 | 1.7994 | 1.7994 | 1.6437 | 1.9113 | 2.0138 | 2.0138 | 2.0138 | 2.0138 |
| 0.05     | 500 | 1.00 | 0.0342 | 0.0342 | 1.5223 | 1.3193 | 1.3193 | 1.3727 | 1.4626 | 1.4626 | 1.4626 | 1.4626 | 1.4626 |
|          | 1000| 1.00 | 0.0122 | 0.0122 | 1.2809 | 1.2809 | 1.2809 | 1.2809 | 1.2809 | 1.2809 | 1.2809 | 1.2809 | 1.2809 |

NOTES: $\delta$ determines the strength of overlap by setting bound $(\delta, 1 - \delta)$ for the true instrument propensity scores. $n$ is the sample size. Mean squared errors (MSE) are all relative to the MSE of the instrumental variables estimator (IV). $|B|$ denotes the average bias in absolute terms over all iterations. Results are based on 10,000 Monte Carlo replications.

Table 6. Design C: relative mean squared errors and absolute biases.

| $\delta$ | $n$ | MSE | IPW1 | IPW2 | BAL1 | BAL2 | BAL3 | BAL4 | eBAL | B2m | B3m |
|----------|-----|-----|------|------|------|------|------|------|------|-----|-----|
| 0.01     | 500 | 1.00 | 4.6478 | 2.8842 | 5700.3 | 2.8842 | 2.8842 | 0.7412 | 0.6223 | 0.8850 | 0.9464 | 0.9595 | 1.0159 |
|          | 1000| 1.00 | 4.6962 | 1.3193 | 1.3193 | 1.3193 | 1.3193 | 1.3193 | 1.3193 | 1.3193 | 1.3193 | 1.3193 | 1.3193 |
| 0.02     | 500 | 1.00 | 3.8383 | 1.7971 | 1.7971 | 1.7971 | 1.7971 | 1.7971 | 1.7971 | 1.7971 | 1.7971 | 1.7971 | 1.7971 |
|          | 1000| 1.00 | 3.8717 | 0.9579 | 0.9579 | 0.9579 | 0.9579 | 0.9579 | 0.9579 | 0.9579 | 0.9579 | 0.9579 | 0.9579 |
| 0.05     | 500 | 1.00 | 2.6347 | 1.4160 | 1.4160 | 1.4160 | 1.4160 | 1.4160 | 1.4160 | 1.4160 | 1.4160 | 1.4160 | 1.4160 |
|          | 1000| 1.00 | 2.6164 | 0.7580 | 0.7580 | 0.7580 | 0.7580 | 0.7580 | 0.7580 | 0.7580 | 0.7580 | 0.7580 | 0.7580 |

NOTES: $\delta$ determines the strength of overlap by setting bound $(\delta, 1 - \delta)$ for the true instrument propensity scores. $n$ is the sample size. Mean squared errors (MSE) are all relative to the MSE of the instrumental variables estimator (IV). $|B|$ denotes the average bias in absolute terms over all iterations. Results are based on 10,000 Monte Carlo replications.

Balancing only covariates and/or selection probabilities (BAL1, BAL2, and BAL3) tends to dominate methods relying on outcome data (BAL4 and eBAL). Using misspecified generated regressors (B2m and B3m) does not hurt the performance of the balancing estimators and can in fact even slightly perform better than their correctly specified counterparts (BAL2 and BAL3) by trading in some bias for further reduction in variance. Unsurprisingly, for all approaches point estimation risk drops with an increase in sample size.

For Design A, the IV estimator serves as a benchmark as it is correctly specified. The IPW2 and all balancing approaches except for BAL2 and B2m are very close in terms of point estimation risk. This is not surprising as under perfect randomization differences in the distributions of observed covariates across instrument levels can only happen by chance. BAL2 and B2m...
perform best as they incorporate exactly the single quantity required for approximate unbiasedness in this constant effect design and do not have inflated variances due to overparameterization of the instrument propensities. For design B, the IV estimator only has a very small bias and thus can serve as a benchmark method. The balancing approaches BAL1, BAL2, and B2m perform better or equal than IPW1 and IPW2 in all cases. The differences are particularly substantial for  and  as in Design B, the IV estimator will be biased due to the nonlinear potential outcome mean. Here, the differences between balancing and MLE based methods are very pronounced. In fact, all balancing approaches lead to a substantial reduction in MSE compared to IV, IPW, and IPW2. In fact, the consistent IPW2 needs a much larger sample size to make up for its larger variance compared to the biased IV estimator. All balancing approaches, however, significantly outperform IV.

In general, the gains over IV are more pronounced for larger samples and larger values of  . All balancing methods also outperform IPW2 by a factor between 1.5 and 5 in terms of MSE. It is interesting that in this design the superior performance of the balancing approaches is often due to a reduction in variance as they sometimes have a slightly larger bias than IPW2. This is not a contradiction to the bias Proposition (5.1), as in this design treatment effects are highly heterogeneous even for the compliers. Therefore, there is no guarantee for a small finite sample bias while the effects on the variance through pushing the inverse probability weights toward more moderate values as outlined in Section 5.2 are still in place. In Design D, IV is severely biased. Note that IPW1, IPW2, BAL2, and B2m here are misspecified due to the quadratic index function in the instrument propensities. BAL4 again has high variance in the

most challenging case ( = 0.01 and  = 500) as in Design B. All other balancing approaches perform extremely well with MSE gains over standard IPW based methods by a factor of 1.8 to 40 most pronounced for small  and . Again BAL2 and B2m lead the pack over BAL1 followed by BAL3 and B3m. The simulation results are evidence that balancing is a superior strategy compared to using likelihood instrument propensity scores for the inverse probability weighting estimator for the LATE in finite samples. A main channel is the reduction in variance, in particular for heterogeneous designs and small sample sizes. This is in line with the literature on balancing weights for selection on observables (Imai and Ratkovic 2014). If one is willing to model the treatment selection process, then including estimates of as balancing variables seems to be beneficial even under some misspecification.

### 8. Empirical Application

This section contains an illustration of the empirical balancing method to the case of one-sided noncompliance. We reevaluate the causal effect of 401(k) retirement plans on private net total financial assets. 401(k) plans are employer-provided, tax-deferred saving plans with partially matched monetary contributions of the employee. The question is whether these plans help to increase net private savings or asset holdings. The evaluation of the effect is complicated by both observed and unobserved individual heterogeneity such as income levels or preferences that might be related to both asset level and the propensity to select the plan. This implies that a simple comparison of savings between participating and non-participating units is likely to yield a biased estimate for the causal effect. We use 401(k) eligibility as a conditionally independent instrument for participation as the former is only provided through the employer. Identification then rests on the assumption that eligibility is independent of possible confounding unobserved

### Table 7. Design D: relative mean squared errors and absolute biases.

|         | IV       | IPW1     | IPW2     | BAL1   | BAL2   | BAL3   | BAL4   | eBAL    | B2m    | B3m    |
|---------|----------|----------|----------|--------|--------|--------|--------|---------|--------|--------|
|  = 0.01 |          |          |          |        |        |        |        |         |        |        |
|  n = 500 | MSE      | 1.0000   | 0.2875   | 2.8227  | 0.0823 | 0.0701 | 0.0922 | 7.7473  | 0.0884 | 0.0680 | 0.0991 |
|         | [B]      | 17.688   | 0.9675   | 4.4397  | 0.5999 | 1.0101 | 0.3608 | 0.0747  | 1.6462 | 1.1449 | 0.3616 |
|  n = 1000 | MSE     | 1.0000   | 0.1244   | 0.3905  | 0.0325 | 0.0265 | 0.0369 | 0.0337  | 0.0403 | 0.0257 | 0.0382 |
|         | [B]     | 17.517   | 2.4996   | 6.5083  | 0.4137 | 0.6889 | 0.3149 | 0.1586  | 0.7646 | 0.8065 | 0.3112 |
|  = 0.02 |          |          |          |        |        |        |        |         |        |        |        |
|  n = 500 | MSE      | 1.0000   | 0.1603   | 1.1849  | 0.0666 | 0.0548 | 0.0690 | 0.0655  | 0.0704 | 0.0532 | 0.0721 |
|         | [B]      | 14.083   | 1.4574   | 4.2963  | 0.3675 | 0.6813 | 0.2569 | 0.0781  | 0.7174 | 0.8016 | 0.2568 |
|  n = 1000 | MSE   | 1.0000   | 0.0915   | 0.2675  | 0.0261 | 0.0222 | 0.0282 | 0.0274  | 0.0335 | 0.0218 | 0.0287 |
|         | [B]     | 14.002   | 2.1236   | 4.8970  | 0.3191 | 0.5419 | 0.2827 | 0.0441  | 0.3219 | 0.6492 | 0.2788 |
|  = 0.05 |          |          |          |        |        |        |        |         |        |        |        |
|  n = 500 | MSE      | 1.0000   | 0.1183   | 0.2399  | 0.0601 | 0.0541 | 0.0650 | 0.0695  | 0.0759 | 0.0536 | 0.0659 |
|         | [B]      | 9.1458   | 0.8476   | 2.2322  | 0.2932 | 0.5165 | 0.2437 | 0.0148  | 0.2551 | 0.6171 | 0.2448 |
|  n = 1000 | MSE   | 1.0000   | 0.0640   | 0.1541  | 0.0311 | 0.0285 | 0.0323 | 0.0339  | 0.0379 | 0.0286 | 0.0326 |
|         | [B]     | 9.1079   | 0.9564   | 2.3517  | 0.2812 | 0.4585 | 0.2617 | 0.0172  | 0.1412 | 0.5515 | 0.2598 |

NOTES:  determines the strength of overlap by setting bound ( ,  − ). for the true instrument propensity scores.  is the sample size. Mean squared errors (MSE) are all relative to the MSE of the instrumental variables estimator (IV). [B] denotes the average bias in absolute terms over all iterations. Results are based on 10,000 Monte Carlo replications.
heterogeneity after conditioning on a set of observed characteristics such as income and other financial and socioeconomic background variables. This and similar identification strategies are widely used in the literature (Benjamin 2003; Abadie 2003; Chernozhukov and Hansen 2004; Chernozhukov et al. 2017), see also Engen, Gale, and Scholz (1996) for a critical assessment.

The data is an excerpt from the Survey of Income and Program Participation (SIPP) of 1991. We apply the same sample restrictions as in Abadie (2003). The final dataset contains 9275 observations. The outcome is measured as total net financial assets in US dollars. The treatment variable is an indicator for participation in a 401(k) plan. The instrument is a binary indicator for eligibility. Observed confounding variables are age, income, family size, education, and other financial and socioeconomic background variables as in Chernozhukov et al. (2017). Note that instead of using the heavily right-skewed income measured in dollars, we use the natural logarithm of income instead to avoid unstable extrapolations for very high-income levels.

We estimate the causal effect of 401(k) participation on net total financial assets using two inverse probability weighting approaches. We also provide standard Wald estimates and replicate the IV result with covariates as in Abadie (2003) for comparison. For the estimators using inverse probability weights, we consider both a simple logistic model and the semiparametric empirical balancing method for estimation of the instrument propensity scores. For both methods, we include all possible interactions of the binary regressors. For the empirical balancing method, we employ an additive spline basis with degree and number of nodes selected via leave-one-out cross-validation using tailored loss function (4.1), that is, the model is chosen to be optimal in terms of its out-of-sample balance. We also experimented with a nonadditive tensor basis but it performed worse in the out-of-sample evaluation.

Figures 1(a) and 1(b) contain the kernel density estimates of the instrument propensity scores for eligible and non-eligible units for both the logistic and the empirical balancing scores. Contrary to the more restrictive logistic model, the semiparametric balancing estimator seems to suggest an almost bimodal distribution for the eligible units with more probability mass around probabilities of 0.65 to 0.75, less between 0.25 and 0.50, and more extreme propensities in the left tail. One can see that for both models there is sufficient overlap in the distribution and minimum scores are sufficiently far away from the boundary. Thus, we do not expect any irregularity problems regarding the identification of the treatment effect parameters (Khan and Tamer 2010; Heiler and Kazak 2021).

Table 8 contains the point estimates of the causal effect and the corresponding standard errors for all methods. The estimates are all significant on a 1% significance level and broadly consistent with the results previously reported in the literature. The Wald estimate of $26,771 differs severely from the other estimates as it is based on the overly restrictive identification assumption of unconditionally independent eligibility. The IV estimates using the model by Abadie (2003) and the parametric IPW estimator yield a very similar result of an around $9500 increase in total net financial assets. The more robust semiparametric balancing estimator suggest about a 30% larger effect. Note that, despite its higher flexibility, empirical balancing has a lower standard error than the more restricted logistic model. This is a consequence of the variance stabilizing effect outlined in Section 5.2. Empirical balancing reduces the presence of extreme propensity scores that can lead to extreme weights which heavily affect the point estimates in finite samples.

9. Concluding Remarks

In this article, we develop estimation methods for the LATE that rely on empirical balancing constraints to improve the internal validity of the causal estimand. The results suggest that there are many theoretical and practical reasons to prefer balancing scores over likelihood based scores for IPW estimation of the

### Table 8. LATE Estimation results.

| Method     | Wald (Abadie) | IV (Likelihood) | IPW (Balancing) |
|------------|---------------|-----------------|-----------------|
| Estimate   | 26711.16      | 9418.83         | 9558.01         |
| Standard Error | 2023.04      | 2152.08         | 1896.57         |

**NOTES:** For Wald and IV, standard errors are calculated using standard heteroscedasticity robust estimators. For both IPW methods, standard errors are obtained via nonparametric bootstrap using 500 replications. All estimates are significant at a 1% significance level.
LATE. As the estimator for the LATE has the typical ratio form, future work should be done to see whether finite sample bias can be further reduced by imposing alternative balancing constraints that do not have to go through two separate reduced form estimates but instead minimize the bias of the ratio directly. An additional contribution would be to derive conditions required that allow to conduct semiparametric inference if generated regressors such as treatment selection probabilities are generated from nonparametric models.

**Supplementary Materials**

The supplementary material contains all proofs and derivations as well as code and replication files.

**Acknowledgments**

I thank to the Editor, three anonymous referees, Michael Knaus, Winfried Pohlmeier, Julian Schissler, Qingyuan Zhao, Toru Kitagawa, and the participants of the Rotterdam Winter Meeting of the Econometric Society 2019 and the St. Gallen Causal Machine Learning Workshop 2020 for fruitful discussions and comments that helped to greatly improve the article. All remaining errors are mine.

**Funding**

The author gratefully acknowledges the financial support from the German Research Foundation through Project 219805061 and by the Graduate School of Decision Sciences (University of Konstanz).

**References**

Abadie, A. (2003), “Semiparametric Instrumental Variable Estimation of Treatment Response Models,” *Journal of Econometrics*, 113, 231–263. [1570,1571,1580]

Angrist, J. D. (1990), “Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security,” *The American Economic Review*, 80, 313–336. [1570]

Angrist, J. D., and Imbens, G. W. (1995), “Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity,” *Journal of the American Statistical Association*, 90, 431–442. [1573]

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996), “Identification of Causal Effects Using Instrumental Variables,” *Journal of the American Statistical Association*, 91, 444–455. [1570,1571]

Angrist, J. D., and Pischke, J.-S. (2009), *Mostly Harmless Econometrics: An Empiricist’s Companion*, Princeton, NJ: Princeton University Press. [1570]

Athey, S., Imbens, G. W., and Wager, S. (2018), “Approximate Residual Balancing: Debiasd Inference of Average Treatment Effects in High Dimensions,” *Journal of the Royal Statistical Series B*, 80, 597–623. [1570]

Benjamin, D. J. (2003), “Does 401(k) Eligibility Increase Saving? Evidence From Propensity Score Subclassification,” *Journal of Public Economics*, 87, 1259–1290. [1580]

Bloom, H. S. (1984), “Accounting for No-shows in Experimental Evaluation Designs,” *Evaluation Review*, 8, 225–246. [1571]

Busso, M., DiNardo, J., and McCrary, J. (2014), “New Evidence on the Finite Sample Properties of Propensity Score Reweighting and Matching Estimators,” *Review of Economics and Statistics*, 96, 885–897. [1572]

Chan, K. C. G., Yam, S. C. P., and Zhang, Z. (2016), “Globally Efficient Non-Parametric Inference of Average Treatment Effects by Empirical Balancing Calibration Weighting,” *Journal of the Royal Statistical Society*, Series B, 78, 673–700. [1574]

Chernozhukov, V., Chetverikov, D., Demirer, M., Dufo, E., Hansen, C., Newey, W., and Robins, J. (2017), “Double/debiased Machine Learning for Treatment and Structural Parameters,” *The Econometrics Journal*, 21, 1–68. [1580]

Chernozhukov, V., and Hansen, C. (2004), “The Effects of 401(k) Participation on the Wealth Distribution: An Instrumental Quatile Regression Analysis,” *Review of Economics and Statistics*, 86, 735–751. [1580]

Donald, S. G., Hsu, Y.-C., and Lieli, R. P. (2014a), “Inverse Probability Weighted Estimation of Local Average Treatment Effects: A Higher Order MSE Expansion,” *Statistics & Probability Letters*, 95, 132–138. [1570,1575]

—— (2014b), “Testing the Unconfoundedness Assumption via Inverse Probability Weighted Estimators of LATE,” *Journal of Business & Economic Statistics*, 32, 395–415. [1570,1574,1575]

Engen, E. M., Gale, W. G., and Scholz, J. K. (1996), “The Illusory Effects of Saving Incentives on Saving,” *Journal of Economic Perspectives*, 10, 113–138. [1580]

Fan, J., Imai, K., Liu, H., Ning, Y., and Yang, X. (2020), “Improving Covariate Balancing Propensity Score: A Doubly Robust and Efficient Approach,” Working Paper. [1570,1573,1575]

Frölich, M. (2007), “Nonparametric IV Estimation of Local Average Treatment Effects with Covariates,” *Journal of Econometrics*, 139, 35–75. [1570,1571,1575]

Graham, B. S., de Xavier Pinto, C. C., and Egel, D. (2012), “Inverse Probability Tilting for Moment Condition Models With Missing Data,” *The Review of Economic Studies*, 79, 1053–1079. [1570,1573]

Hahn, J. (1998), “On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects,” *Econometrica*, 66, 315–331. [1570]

Hainmueller, J. (2012), “Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies,” *Political Analysis*, 20, 25–46. [1570]

Heckman, J. J., and Vytlacil, E. J. (2005), “Structural Equations, Treatment Effects, and Econometric Policy Evaluation,” *Econometrica*, 73, 669–738. [1576]

Heiler, P., and Kazak, E. (2021), “Valid Inference for Treatment Effect Parameters Under Irregular Identification and Many Extreme Propensity Scores,” *Journal of Econometrics*, 222, 1083–1108. [1571,1580]

Heiler, P., and Mareckova, J. (2021), “Shrinkage for Categorical Regressors,” *Journal of Econometrics*, 223, 161–189. [1576]

 Hirano, K., Imbens, G. W., and Ridder, G. (2003), “Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score,” *Econometrica*, 71, 1161–1189. [1570,1574,1575]

Hong, H., and Nekipelov, D. (2010), “Semiparametric Efficiency in Nonlinear LATE Models,” *Quantitative Economics*, 1, 279–304. [1575]

Imai, K., and Ratkovic, M. (2014), “Covariate Balancing Propensity Score,” *Journal of the Royal Statistical Society*, Series B, 76, 243–263. [1570,1572,1579]

Imbens, G. W., and Angrist, J. D. (1994), “Identification and Estimation of Local Average Treatment Effects,” *Econometrica*, 62, 467–475. [1571]

Khan, S., and Tamer, E. (2010), “Irregular Identification, Support Conditions, and Inverse Weight Estimation,” *Econometrica*, 78, 2021–2042. [1580]

Knaack, M. C., Lechner, M., and Reimers, A. K. (2020), “For Better or Worse. The Effects of Physical Education on Child Development,” *Labour Economics*, 67, 101904. [1570]

Li, F., Morgan, K. L., and Zaslavsky, A. M. (2018), “Balancing Covariates Via Propensity Score Weighting,” *Journal of the American Statistical Association*, 113, 390–400. [1574]

Li, Q., Racine, J. S., and Wooldridge, J. M. (2009), “Efficient Estimation of Average Treatment Effects with Mixed Categorical and Continuous Data,” *Journal of Business & Economic Statistics*, 27, 206–223. [1575]

Ning, Y., Sida, P., and Imai, K. (2020), “Robust Estimation of Causal Effects Via a High-Dimensional Covariate Balancing Propensity Score,” *Biometrika*, 107, 533–554. [1570]

Ouyang, D., Li, Q., and Racine, J. S. (2009), “Nonparametric Estimation of Regression Functions With Discrete Regressors,” *Econometric Theory*, 25, 1–42. [1576]

Pohlmeier, W., Seiberlich, R., and Uysal, S. D. (2016), “A Simple and Successful Shrinkage Method for Weighting Estimators of Treatment Effects,” *Computational Statistics & Data Analysis*, 100, 512–525. [1572]

Peterba, J. M., Yenti, S. F., and Wise, D. A. (1995), “Do 401(k) Contributions Crowd out other Personal Saving?” *Journal of Public Economics*, 58, 1–32. [1570]
Robins, J. M., and Rotnitzky, A. (1995), “Semiparametric Efficiency in Multivariate Regression Models with Missing Data,” Journal of the American Statistical Association, 90, 122–129. [1573]
Rubin, D. B. (2007), “The Design Versus the Analysis of Observational Studies for Causal Effects: Parallels with the Design of Randomized Trials,” Statistics in Medicine, 26, 20–36. [1569,1571,1575]
SantAnna, P. H., and Zhao, J. (2020), “Doubly Robust Difference-in-Differences Estimators,” Journal of Econometrics, 219, 101–122. [1573]
Savage, L. J. (1971), “Elicitation of Personal Probabilities and Expectations,” Journal of the American Statistical Association, 66, 783–801. [1572]
Vytlačil, E. (2002), “Independence, Monotonicity, and Latent Index Models: An Equivalence Result,” Econometrica, 70, 331–341. [1575]
Zhao, Q. (2019), “Covariate Balancing Propensity Score by Tailored Loss Functions,” The Annals of Statistics, 47, 965–993. [1570,1574]
Zhao, Q., and Percival, D. (2017), “Entropy Balancing is Doubly Robust,” Journal of Causal Inference, 5. [1570]
Zubizarreta, J. R. (2015), “Stable Weights That Balance Covariates for Estimation with Incomplete Outcome Data,” Journal of the American Statistical Association, 110, 910–922. [1570]