Center for Evaluation and Development
WORKING PAPER SERIES

Endogeneity and Non-Response Bias in Treatment Evaluation: Nonparametric Identification of Causal Effects by Instruments

Working Paper 2015/5

Hans Fricke
Markus Frölich
Martin Huber
Michael Lechner

ABSTRACT

This paper proposes a nonparametric method for evaluating treatment effects in the presence of both treatment endogeneity and attrition/non-response bias, using two instrumental variables. Making use of a discrete instrument for the treatment and a continuous instrument for non-response/attrition, we identify the average treatment effect on compliers as well as the total population and suggest non- and semiparametric estimators. We apply the latter to a randomized experiment at a Swiss University in order to estimate the effect of gym training on students' self-assessed health. The treatment (gym training) and attrition are instrumented by randomized cash incentives paid out conditional on gym visits and by a cash lottery for participating in the follow-up survey, respectively.

JEL Classification: C14, C21, C23, C24, C26

Keywords: local average treatment effect, attrition, endogeneity, weighting, instrument, experiment

Corresponding author:
Markus Frölich
University of Mannheim
L7, 3-5
68131 Mannheim, Germany
E-Mail: froelich@uni-mannheim.de

1 Introduction

The evaluation of the causal effect of some treatment, e.g. a health or labor market intervention, on an outcome variable, e.g. individual health or labor market performance, is frequently complicated by two identification problems: (i) endogeneity due to non-random selection into treatment and (ii) non-response/attrition, e.g. selective non-response with respect to the follow-up survey in which the outcome is measured. The methodological contribution of this paper is to suggest a nonparametric approach that tackles either problem based on two distinct instruments in order to identify average treatment effects. To the best of our knowledge, this is the first paper to develop fully nonparametric identification results to solve the endogeneity and attrition issues via instrumental variables.

The main identification result focuses on the case of a binary treatment and a binary instrument for the treatment, which fits the framework of social experiments with non-compliance, where randomization of the treatment serves as instrument and actual take-up as treatment. However, in analogy to the discussion in Frölich (2007) (who considers the case of treatment endogeneity without attrition), the findings can be generalized to multi-valued instruments. Concerning endogenous outcome non-response, we assume the respective instrument to be continuous. Financial incentives for responding to a follow-up survey, see e.g. Castiglioni, Pforr, and Krieger (2008), and/or the number of phone calls when contacting potential survey participants, see Behaghel, Crépon, Gurgand, and Le Barbanchon (2012), may for instance serve as instruments, if their support is sufficiently rich and the IV assumptions appear plausible in the empirical context. We show that our assumptions allow identifying the local average treatment effect (LATE) among compliers as well as the average treatment effect (ATE) and suggest nonparametric estimation approaches based on regression and weighting. We also provide a simulation study that suggests that non- and semiparametric versions of the regression-based estimators perform well in samples with several 1000 observations, which is quite common in recent social experiments.

We apply our methods to a social experiment conducted at the University of St. Gallen in Switzerland to assess the effect of students' physical (gym) training on health. The application is rather unique in the sense that it contains two separately randomized (and thus, highly credible) instruments for both the treatment and non-response. Firstly, the treatment of interest, training in the university's gym facilities, is instrumented by a randomized cash incentive (100 CHF)

paid out conditional on actual gym visits measured by a scanner system. Secondly, attrition is instrumented by a cash lottery for participating in the follow-up survey in which the outcome is measured. Cash was only paid out conditional on answering the survey. Importantly, the amount offered for participating in the survey was randomly varied between 0 and 200 CHF, so that the instrument is (quasi-)continuous. We observe that this cash incentive and its amount had a strong effect on response behavior. On the other hand, for the treatment of interest, physical training, we do not find any significant short run effects on self-assessed health.

This paper adds to the treatment evaluation literature by considering both treatment endogeneity and outcome attrition, as a brief review of previous studies demonstrates. The seminal papers of Imbens and Angrist (1994) and Angrist, Imbens, and Rubin (1996) discuss LATE identification under an unconditionally valid instrument for the treatment, while Abadie (2003), Frölich (2007), and Tan (2006) propose semi- and nonparametric approaches when the IV assumptions only hold conditional on observed characteristics. However, none of these studies consider the problem of outcome non-response or attrition. The latter is frequently modelled by a so-called missing at random (MAR) restriction, which imposes the conditional exogeneity (or independence with respect to potential outcomes) of attrition given observed characteristics, see for instance Rubin (1976), Little and Rubin (1987), Robins, Rotnitzky, and Zhao (1994), Fitzgerald, Gottschalk, and Moffitt (1998), and Abowd, Crepon, and Kramarz (2001), among many others. As an alternative to MAR which is suitable for the LATE framework, Frangakis and Rubin (1999) propose their so-called latent ignorability (LI) assumption. The latter requires that attrition is exogenous conditional on the treatment compliance behavior, which characterizes how an individual's treatment status reacts on its instrument, and possibly further observed covariates. See also Mealli, Imbens, Ferro, and Biggeri (2004) for related LI assumptions and Frölich and Huber (2014) for LATE estimation under MAR and LI in dynamic attrition models with multiple outcome periods.

A shortcoming of LI (and MAR) is that attrition must not be related in a very general way to unobservables affecting the outcome, whereas our approach allows for such non-ignorable non-response, attrition, or sample selection through the availability of a distinct instrument for non-response/attrition. The early work on non-ignorable non-response models imposed rather strong parametric assumptions, see for instance Heckman (1976, 1979) and Hausman and Wise (1979), which entail identification through their tight functional form restrictions.

Instruments for attrition may, however, serve as additional source of identification and help preventing multicollinearity problems. Instruments for attrition are imperative in nonparametric models with non-ignorable non-response, see for instance Huber (2012, 2014), who focusses on attrition problems in treatment evaluation with an exogenous treatment, at least conditional on observables. Zhang, Rubin, and Mealli (2009) and Frumento, Mealli, Pacini, and Rubin (2012) evaluate treatment effects under both endogeneity and non-ignorable non-response, but assume that there only exists a valid instrument for the treatment. Identification therefore relies on tight parametric assumptions, which need not be imposed here.[1]

In the application, we exploit a unique dataset where both treatment eligibility and response incentives were randomized. While many studies assess the LATE in social experiments by using treatment randomization as instrument for actual treatment take-up, instruments for non-ignorable attrition are rarely considered. One exception is DiNardo, McCrary, and Sanbonmatsu (2006), who apply the parametric estimator suggested by Heckman (1976, 1979) and use the effort to interview study subjects as instrument for attrition in a randomized trial of the 'Moving to Opportunity' program. Furthermore, Behaghel, Crépon, Gurgand, and Le Barbanchon (2012) use phone calls as (quasi-)instrument for non-response to derive bounds on the treatment effect in a French job search experiment. However, neither of these studies consider the problem of treatment non-compliance. Furthermore, both studies assume a discrete (rather than continuous) instrument for attrition, so that point identification is only obtained under strong parametric restrictions (see DiNardo, McCrary, and Sanbonmatsu (2006)), while more flexible (nonparametric) modelling only allows for a partial identification of the treatment effect (see Behaghel, Crépon, Gurgand, and Le Barbanchon (2012)). We are not aware of any other empirical study that is based on two different randomized instruments for tackling both treatment endogeneity and outcome attrition in a nonparametric model. Using the proposed methods will permit designing future experimental and nonexperimental studies that retain their validity despite issues of selective non-response and non-compliance.

The remainder of this paper is organized as follows. Section 2 introduces a nonparametric treatment effect model with endogeneity and outcome attrition. Section 3 introduces the IV assumptions and develops the nonparametric identification approaches, and Section 4 discusses estimation. Section 5 presents simulation evidence on the finite sample properties of the estimation

[1] See also Semykina and Wooldridge (2010) and Schwiebert (2012) for further semiparametric models with endogeneity and selection bias.

approach. Section 6 discusses the application to the sports experiment at the University of St. Gallen. Section 7 concludes.

2 Model

Assume that we would like to evaluate the effect of a binary treatment D on an outcome variable Y. The latter is, however, only partially observed conditional on response, measured by the binary indicator R. Furthermore, we observe a vector of baseline covariates, denoted by X. Identification will be based on two instruments Z_1 and Z_2 for the endogenous treatment and the non-ignorable non-response. To this end, we postulate the following structural model consisting of a nonparametric system of equations characterizing the outcome, response, and the treatment:

$$Y = \varphi(D, X) + U \tag{1}$$

$$R = 1\,(\ \zeta(D, Z_2, X) \geq V\ ) \tag{2}$$

$$D = 1\,(\ \chi(Z_1, X) \geq W\ )\,. \tag{3}$$

$$Y \text{ is observed only when } R = 1$$

φ, ζ, χ denote unknown functions so that our model is fully nonparametric. $1\,(\cdot)$ is the indicator function which is equal to one if its argument is true and zero otherwise. U, V, W are unobservables and may be arbitrarily associated, so that the treatment is in general endogenous and non-response is non-ignorable. That is, both the treatment and non-response are related to unobservables that affect the outcome. The elements of X are not required to be exogenous either, but may be related to the unobservables, as long as the identifying assumptions discussed further below hold. Z_1 denotes the instrument for treatment D, henceforth referred to as first instrument. Z_1 is assumed to be *binary* for the ease of exposition, even though the discussion could be extended to multi-valued instruments with bounded support, see also Frölich (2007). Z_2 is the instrument for response R, henceforth referred to as second instrument, which is assumed to be *continuous*. In our model, we permit the two instruments to be possibly correlated and the compliance type to be correlated with Z_2.

Our model imposes additive separability in the outcome equation (1). This structure implicitly invokes a conditional constant-treatment effect, i.e. that the treatment effect is identical for all individuals with same X, see also Angrist and Fernández-Val (2010). On the other hand, it permits for arbitrary heterogeneity across X since the function φ is completely unrestricted. Hence,

by including more covariates in X we can enrich the amount of heterogeneity permitted. The advantage of the additive separability in the outcome equation (1) is that we can weaken the support requirements on the instrument Z_2. The conventional approach to tackle non-response in (nonseparable) nonparametric models is to assume that the instrument Z_2 is so strong that, for every value of V, it can make people respond, such that for every $V = v$ (and for every X and Z_1) the response probability is positive.[2] However, in many applications, including ours, the instrument Z_2 is not strong enough to believe that it makes everyone respond. This applies to all empirical problems where non-response does not vanish completely for some value of z_2. Therefore, we want to permit that the outcome is never observed for a range of values of V, due to Z_2 not being sufficiently strong to affect the response behavior.

Our model can be easily translated into the potential outcome notation, see for instance Rubin (1974). Let Y^d, R^d denote the potential outcome and the potential response state under treatment $d \in \{0, 1\}$, i.e. when exogenously setting the treatment to either state. For an individual i in the population, these parameters are defined as follows under our model:

$$\begin{aligned} R_i^d &= 1\left(\zeta(d, Z_{2i}, X_i) \geq V_i \right), \\ Y_i^d &= \varphi(d, X_i) + U_i. \end{aligned}$$

Hence, we permit that the treatment D not only affects the outcomes but also the response behavior. Estimation of the treatment effects is thus complicated through two channels: First, V and U might be correlated with each other as well as with W. Second, through $\zeta(d, Z_2, X)$ the treatment D itself has an impact on whose outcomes are observed. Similarly, we define the potential treatment states as a function of the first instrument, i.e. for $z_1 \in \{0, 1\}$,

$$D_i(z_1) = 1\left(\chi(z_1, X_i) \geq W_i \right).$$

As discussed in Angrist, Imbens, and Rubin (1996), the population can be categorized into four compliance types (denoted by T), according to the treatment behavior as a function of the first instrument: The *always takers* ($T_i = a$) take treatment irrespective of Z_1, i.e. $D_i(0) = D_i(1) = 1$. The *never takers* ($T_i = n$) do not take the treatment irrespective of Z_1, i.e. $D_i(0) = D_i(1) = 0$. The *compliers* ($T_i = c$) take the treatment only if Z_1 is one, i.e. $D_i(0) = 0, D_i(1) = 1$. Finally, the *defiers* ($T_i = d$) take the treatment only if Z_1 is zero, i.e. $D_i(0) = 1, D_i(1) = 0$.

[2] This has often been referred to as 'identification at infinity' in the parametric literature on attrition, non-response and selection models.

In the absence of non-response, Imbens and Angrist (1994) showed the identification of the local average treatment effect (LATE) on the compliers, i.e. $E[Y^1 - Y^0|\mathcal{T} = c]$ under the assumptions that Z_1 is independent of the potential outcomes and treatment states and defiers do not exist (i.e. weak monotonicity of D in Z_1). Abadie (2003), Frölich (2007), Tan (2006) relax the IV assumptions to only hold conditional on X. In this paper, we in addition permit for attrition and non-response, which generally entails selection bias through associations of V with U and/or W and therefore motivates the use of the second instrument Z_2.

3 Identification

3.1 Assumptions and main identification results

This section discusses our IV assumptions and shows the identification of the LATE and the ATE. The first assumption requires the instruments to be independent of the unobservables U, V, W conditional on X, which may itself be endogenous (i.e. confounded by the unobservables). While Abadie (2003), Frölich (2007), and Tan (2006) invoke a similar assumption for Z_1 only, conditional independence needs to hold for both instruments Z_1 and Z_2 in our model with endogeneity and attrition. For the ease of exposition, Assumption 1 is slightly stronger than needed for the various results to follow. We express the independence condition with respect to type T and not with respect to the unobservable W, as we later only require independence within the types and not for each value of W.

Assumption 1: IV independence

$$Z_1 \perp\!\!\!\perp T|X, Z_2$$

$$(Z_1, Z_2) \perp\!\!\!\perp (U, V)|X, T$$

where the symbol $\perp\!\!\!\perp$ denotes statistical independence. It is worth noting that Assumption 1 would be implied e.g. by the following stronger assumption:

$$(Z_1, Z_2) \perp\!\!\!\perp (U, V, W)|X. \tag{4}$$

The main difference is that Assumption 1 permits Z_2 and W to be dependent, whereas (4) does not. As W determines the type, i.e. whether someone is a complier, always taker, or never taker, permitting dependence between Z_2 and W could be relevant in applications where Z_2 is not fully

randomly assigned but possibly dependent on treatment choice. Assumption 1 also allows for an association between Z_1 and W, as long as the dependence vanishes when conditioning on Z_2.

The stronger assumption (4) is not required for the identification results. If it is nevertheless imposed, for instance because both instruments are randomized independently of each other as in our application, it implies that the probability of being a complier does not depend on Z_2. This condition is testable because $\Pr(T = c|Z_2, X)$, i.e. the proportion of compliers given Z_2 and X, is identified further below. It would further imply $Z_2 \perp\!\!\!\perp D|X, Z_1$. Hence, in applications where both assumptions appear equally plausible, this may be used to construct partial tests for identification. One could strengthen assumption (4) even further by assuming that the X variables are also exogenous, i.e. independent of the unobservables. This could help to increase the identification region particularly if the common support assumption discussed further below is not satisfied in an application.

Assumption 1 implies that the first instrument is conditionally independent of the potential treatment states $D(1), D(0)$ and does not have a direct effect on response behavior or the outcome through V or U. Z_1 may for instance be the assignment indicator in a randomized experiment. The potential treatment states are independent of Z_1 under a successful randomization and the independence of Z_1 and (U, V) is satisfied if the random assignment itself does not affect R and Y other than through D. In observational studies, on the other hand, (but also in experiments where randomization is within strata defined on X), Assumption 1 is often only plausible after conditioning on covariates X.

In addition to the independence assumptions, identification requires a monotonicity condition. Assumption 2 imposes weak monotonicity of the treatment in its instrument, which rules out the existence of defiers, and further invokes the existence of compliers.[3]

Assumption 2: Weak monotonicity of treatment choice

$$\begin{aligned} \Pr(T = c) &> 0 \\ \Pr(T = d) &= 0. \end{aligned}$$

[3] Alternatively, one could also impose weakly negative monotonicity (allowing for defiers, but ruling out compliers). As both cases are symmetric, we only consider weakly positive monotonicity in the remainder of the paper. Note further that part (i) of the assumption is directly testable. In contrast, part (ii) is mostly untestable (although see Huber and Mellace (2015) and Kitagawa (2015) for recent methods jointly testing monotonicity *and* IV independence).

Next, we define $\pi(x) = \Pr(Z_1 = 1|X = x)$ and $p(z_2, x) = \Pr(Z_1 = 1|X = x, Z_2 = z_2)$ and the corresponding random variables, i.e. for random X and Z_2, as

$$\begin{aligned} \Pi &= \pi(X) = \Pr(Z_1 = 1|X), \\ P &= p(Z_2, X) = \Pr(Z_1 = 1|Z_2, X). \end{aligned}$$

Our third assumption states that the two probabilities P and Π need to be different from 0 and 1. This common support restriction implies that for every value of z_2 and x, observations with both $Z_1 = 0$ and $Z_1 = 1$ exist.

Assumption 3: Variation of the instruments

$$0 < \Pi < 1, \quad 0 < P < 1.$$

It follows from Assumptions 1 and 3 that the fraction of compliers is identified as

$$\Pr(T = c) = E\left[\frac{D}{P}\frac{Z_1 - P}{1 - P}\right]. \tag{5}$$

As a further definition, let

$$\psi_d(z_2, x) \equiv \frac{E[R(Z_1 - p(z_2, x))|D = d, X = x, Z_2 = z_2]}{E[Z_1 - p(z_2, x)|D = d, X = x, Z_2 = z_2]}, \tag{6}$$

for $d \in \{0, 1\}$, and define the corresponding random variable for random Z_2 and X as

$$\Psi_d = \psi_d(Z_2, X).$$

Under our previous assumptions we can derive the following lemma:

Lemma 1: Under Assumptions 1 to 3 the conditional distribution functions of V are identified as:

$$\begin{aligned} \psi_1(z_2, x) &= F_{V|X=x,T=c}(\zeta(1, z_2, x)), \\ \psi_0(z_2, x) &= F_{V|X=x,T=c}(\zeta(0, z_2, x)). \end{aligned}$$

Since the left hand side is identified, see definition (6), the distribution function of V at different values of ζ is identified, too.

Our identification strategy also requires the unobservable V to be continuously distributed, which appears rather natural in most applications and motivates Assumption 4:

Assumption 4: The distribution function $F_{V|X,T=c}(v)$ is strictly increasing in v.

By combining Lemma 1 with Assumption 4 we obtain that if some values x, z_2', z_2'' satisfy $\psi_1(z_2', x) = \psi_0(z_2'', x)$, then we also have that $\zeta(1, z_2', x) = \zeta(0, z_2'', x)$. This result will be crucial for identification, which is based on the following intuition. Note that the sample selection/non-response problem occurs because we observe outcomes only for observations with $V \leq \zeta(D, Z_2, X)$. The sets of values V which satisfy this condition differ for $D = 0$ and $D = 1$. At the same time, D and V are correlated. If we can find values of the instrument such that $\zeta(1, z_2', x) = \zeta(0, z_2'', x) = a$, then the set of observations with outcome data is given by $V \leq a$ in the treated and non-treated population.

For a more formal illustration, consider the following expression for some value x and z_2':

$$E\left[YRD|X = x, Z_2 = z_2', Z_1 = 1\right] - E\left[YRD|X = x, Z_2 = z_2', Z_1 = 0\right]. \tag{7}$$

Via partitioning each expression by the three types (a, c, n) one can show that (7) equals

$$= E\left[\{\varphi(1, x) + U\} \cdot 1\left\{\ \zeta(1, z_2', x) \geq V\ \right\} | X = x, T = c\right] \Pr\left(T = c|X = x, Z_2 = z_2'\right),$$

where we used $(U, V) \perp\!\!\!\perp (Z_1, Z_2)|X, T$. Similarly, for some value x and z_2''

$$\begin{aligned} & E\left[YR(1-D)|X = x, Z_2 = z_2'', Z_1 = 1\right] - E\left[YR(1-D)|X = x, Z_2 = z_2'', Z_1 = 0\right] \\ = \ & -E\left[\{\varphi(0, x) + U\} \cdot 1\left\{\ \zeta(0, z_2'', x) \geq V\ \right\} | X = x, T = c\right] \Pr\left(T = c|X = x, Z_2 = z_2''\right). \end{aligned}$$

Using the results in the appendix, the previous expressions may (after some tedious calculations) be reformulated as follows:

$$\begin{aligned} & \frac{E\left[YR(Z_1 - p(z_2', x))|D = 1, X = x, Z_2 = z_2'\right]}{E\left[Z_1 - p(z_2', x)|D = 1, X = x, Z_2 = z_2'\right]} - \frac{E\left[YR(Z_1 - p(z_2'', x))|D = 0, X = x, Z_2 = z_2''\right]}{E\left[Z_1 - p(z_2'', x)|D = 0, X = x, Z_2 = z_2''\right]} \quad (8) \\ & = E\left[\{\varphi(1, x) + U\} \cdot 1\left\{\ \zeta(1, z_2', x) \geq V\ \right\} - \{\varphi(0, x) + U\} \cdot 1\left\{\ \zeta(0, z_2'', x) \geq V\ \right\} | X = x, T = c\right] \end{aligned}$$

Now suppose the values z_2' and z_2'' are chosen such that $\psi_1(z_2', x) = \psi_0(z_2'', x)$. Assumption 4 implies that $F_{V|X,T=c}$ is invertible or, in other words, that if $\psi_1(z_2', x) = \psi_0(z_2'', x)$ it also holds that $\zeta(1, z_2', x) = \zeta(0, z_2'', x)$ by Lemma 1. Hence,

$$\begin{aligned} & = E\left[\{\varphi(1, x) + U - \varphi(0, x) - U\} \cdot 1\left\{\ \zeta(1, z_2', x) \geq V\ \right\} | X = x, T = c\right] \\ & = \{\varphi(1, x) - \varphi(0, x)\} \cdot E\left[1\left\{\ \zeta(1, z_2', x) \geq V\ \right\} | X = x, T = c\right] \\ & = \{\varphi(1, x) - \varphi(0, x)\} \cdot \psi_1(z_2', x). \end{aligned}$$

Therefore, the following expression identifies the treatment effect *conditional* on X:

$$E[Y^1 - Y^0|X = x, T = c] = \\ \frac{1}{\psi_1(z_2', x)} \left[\frac{E\left[YR(Z_1 - p(z_2', x))|D = 1, X = x, Z_2 = z_2'\right]}{E\left[Z_1 - p(z_2', x)|D = 1, X = x, Z_2 = z_2'\right]} - \frac{E\left[YR(Z_1 - p(z_2'', x))|D = 0, X = x, Z_2 = z_2''\right]}{E\left[Z_1 - p(z_2'', x)|D = 0, X = x, Z_2 = z_2''\right]} \right]. \tag{9}$$

For obtaining the LATE we need to identify $E[Y^1 - Y^0|X, T = c]$ at almost every x in the complier population. This requires that for every x some values z_2' and z_2'' exist that satisfy $\psi_1(z_2', x) = \psi_0(z_2'', x)$. Let $Supp\left(\Psi_1|X = x\right)$ denote the support of Ψ_1 in the $X = x$ subpopulation and analogously for Ψ_0. Furthermore, denote the common support conditional on x as

$$\mathcal{X}_x \equiv Supp\left(\Psi_1|X = x\right) \cap Supp\left(\Psi_0|X = x\right). \tag{10}$$

If, for some value x, the common support $\mathcal{X}_x$ is non-empty, there is at least one pair of values z_2', z_2'' that satisfies $\psi_1(z_2', x) = \psi_0(z_2'', x)$. We impose the following common support restriction.

Assumption 5: For almost every x (in the complier population), the common support $\mathcal{X}_x$ is non-empty.

Assumption 5 guarantees that LATE is identified since the conditional treatment effect is identified almost everywhere.

For every x, it is in principle sufficient if we just pick one point of $\mathcal{X}_x$ and apply (9). However, for the sake of sufficient precision in estimation, we would rather prefer to make use of all values contained in $\mathcal{X}_x$. As shown in the appendix, we can also identify (9) via conditioning on Ψ_d instead of Z_2. Let $\eta \in \mathcal{X}_x$ be some value from the common support. One can show that

$$E\left[Y^1 - Y^0|X = x, T = c\right] = \frac{1}{\eta}\left(\ \Xi_1(x, \eta) - \Xi_0(x, \eta)\ \right)$$

where

$$\Xi_d(x, \eta) = \frac{E\left[\frac{YR}{E[Z_1|Z_2, X=x, \Psi_d=\eta]} \frac{Z_1 - E[Z_1|Z_2, X=x, \Psi_d=\eta]}{1 - E[Z_1|Z_2, X=x, \Psi_d=\eta]} | D = d, X = x, \Psi_d = \eta\right]}{E\left[\frac{1}{E[Z_1|Z_2, X=x, \Psi_d=\eta]} \frac{Z_1 - E[Z_1|Z_2, X=x, \Psi_d=\eta]}{1 - E[Z_1|Z_2, X=x, \Psi_d=\eta]} | D = d, X = x, \Psi_d = \eta\right]}. \tag{11}$$

Since the previous result holds for any η, we could exploit all the information available in the data by taking an average over all values $\eta \in \mathcal{X}_x \subseteq [0, 1]$ for a given x. Consider any arbitrary weighting function $w(\eta, x)$ as a function of η and possibly also of x. The conditional treatment effect is given

by

$$E\left[Y^1 - Y^0 | X = x, T = c\right] = \frac{\int_0^1 \frac{1}{\eta}\left(\ \Xi_1(x,\eta) - \Xi_0(x,\eta)\ \right) w(\eta, x) d\eta}{\int_0^1 w(\eta, x) d\eta}, \tag{12}$$

provided that the weighting function $w(\eta, x)$ does not integrate to zero. Therefore, integration over X gives the LATE. Since the additive separability of the outcome equation (1) implies that $E\left[Y^1 - Y^0 | X = x, T = c\right] = E\left[Y^1 - Y^0 | X = x\right]$ (similarly as in Angrist and Fernández-Val (2010)), also the ATE is obtained in an analogous way. As can be seen from our main identification result presented in Theorem 1, the LATE and ATE only differ in terms of the weighting of the covariates X. To ease notation, we denote the integral of the weight function for a value of x as

$$c(x) = \int w(\eta, x) d\eta, \tag{13}$$

and we suppose that $c(x)$ is non-zero for almost every x.

Theorem 1: Under Assumptions 1 to 5 we obtain

$$E\left[Y^1 - Y^0 | T = c\right] = \frac{1}{E\left[\frac{D}{P}\frac{Z_1 - P}{1-P}\right]} \int \int_0^1 \left(\ \Xi_1(X,\eta) - \Xi_0(X,\eta)\ \right) \frac{w(\eta, X)}{\eta \cdot c(X)} E\left[\frac{D}{P}\frac{Z_1 - P}{1-P} | X\right] d\eta dF_X$$

and (14)

$$E\left[Y^1 - Y^0\right] = \int \int_0^1 \left(\ \Xi_1(X,\eta) - \Xi_0(X,\eta)\ \right) \frac{w(\eta, X)}{\eta \cdot c(X)} d\eta dF_X. \tag{15}$$

3.2 Identification results for independent instruments

In our application, the second instrument Z_2 is randomized independently of Z_1. This has two implications, which lead to considerable simplifications of the previous formulae. First, the fraction of compliers is independent of Z_2, i.e. $\Pr(T = c|X, Z_2) = \Pr(T = c|X) = \Pr(T = c|X, \Psi_1)$, where the last equality follows because $\Psi_1 = \psi_1(Z_2, X)$ is only a function of Z_2 and X. Second, Z_1 and Z_2 are independent such that $\Pr\left(Z_1 = 1|Z_2, X, \Psi_1\right) = \Pr(Z_1 = 1|Z_2, X) = \Pr(Z_1 = 1|X)$ and therefore, $P = \Pi$. This also implies that D is independent of Ψ_d given X.[4] The control function

[4]Proof: $E\left[D|\Psi_d, X\right] = E\left[D|\Psi_d, X, T = c\right] \Pr\left(T = c|\Psi_d, X\right)$ $+E\left[D|\Psi_d, X, T = a\right] \Pr\left(T = a|\Psi_d, X\right) = E\left[Z_1|\Psi_d, X, T = c\right] \Pr\left(T = c|\Psi_d, X\right) + \Pr\left(T = a|\Psi_d, X\right) = E\left[Z_1|\Psi_d, X, T = c\right] \Pr\left(T = c|X\right) + \Pr\left(T = a|X\right)$. Using the independence of Z_1 and Z_2 we obtain $= E\left[Z_1|X, T = c\right] \Pr\left(T = c|X\right) + \Pr\left(T = a|X\right)$, which completes the proof.

thus simplifies to

$$\psi_d(z_2, x) \equiv \frac{E\left[R\left(Z_1 - \pi(x)\right)|D = d, X = x, Z_2 = z_2\right]}{E\left[Z_1 - \pi(x)|D = d, X = x, Z_2 = z_2\right]}. \tag{16}$$

We also note that $E\left[\frac{D}{\Pi}\frac{Z_1-\Pi}{1-\Pi}|X = x, \Psi_d = \eta\right] = E\left[\frac{D}{\Pi}\frac{Z_1-\Pi}{1-\Pi}|X = x\right] = \Pr\left(T = c|X = x\right)$ and that the expressions of Theorem 1 simplify considerably, see Lemma 2. In addition, we can also express the treatment effects based on a weighting expression, which bears some similarities to inverse probability weighting (IPW, see the seminal work of Horvitz and Thompson (1952)), as one can show by using iterated expectations.

Lemma 2: Under Assumptions 1 to 5 and $Z_2 \perp\!\!\!\perp (Z_1, T)|X$ the average treatment effects are identified as

$$\begin{aligned}
& E\left[Y^1 - Y^0|T = c\right] \\
= \; & \frac{1}{\Pr(T = c)} \int_0^1 \int_0^1 \left(E\left[\frac{YRD}{\Pi}\frac{Z_1 - \Pi}{1 - \Pi}|X, \Psi_1 = \eta\right] + E\left[\frac{YR(1-D)}{\Pi}\frac{Z_1 - \Pi}{1 - \Pi}|X, \Psi_0 = \eta\right]\right) \frac{w(\eta, X)}{\eta \cdot c(X)} d\eta dF_X \\
= \; & \frac{1}{\Pr(T = c)} E\left[YR\frac{Z_1 - \Pi}{\Pi\left(1 - \Pi\right) \cdot c(X)} \cdot \left\{D\frac{w(\Psi_1, X)}{\Psi_1 \cdot f\left(\Psi_1|X\right)} + (1 - D)\frac{w(\Psi_0, X)}{\Psi_0 \cdot f\left(\Psi_0|X\right)}\right\}\right].
\end{aligned} \tag{17}$$

and

$$\begin{aligned}
& E\left[Y^1 - Y^0\right] \\
= \; & \int_0^1 \int_0^1 \left(\frac{E\left[\frac{YRD}{\Pi}\frac{Z_1-\Pi}{1-\Pi}|X, \Psi_1 = \eta\right] + E\left[\frac{YR(1-D)}{\Pi}\frac{Z_1-\Pi}{1-\Pi}|X, \Psi_0 = \eta\right]}{\Pr\left(T = c|X = x\right)} \right) \frac{w(\eta, X)}{\eta \cdot c(X)} d\eta dF_X \\
= \; & E\left[\frac{YR}{\Pr(T = c|X)}\frac{Z_1 - \Pi}{\Pi\left(1 - \Pi\right) \cdot c(X)} \cdot \left\{D\frac{w(\Psi_1, X)}{\Psi_1 \cdot f\left(\Psi_1|X\right)} + (1 - D)\frac{w(\Psi_0, X)}{\Psi_0 \cdot f\left(\Psi_0|X\right)}\right\}\right]
\end{aligned} \tag{18}$$

Natural estimators follow from replacing unconditional expectations in the latter equations by sample means and plugging in nonparametric estimators of the other components, see equations (23), (24), (25), and (26) in Section 4. Under standard regularity conditions, the estimators proposed further below are consistent for any non-zero weighting function w. Specifically, it is required that π is bounded away from zero and one and f is bounded away from zero for all values of ψ_d and x with a non-zero $w(\psi_d, x)$. As the treatment effect is identified for any (non-zero) weighting function, the latter should ideally be chosen such that it minimizes the variance of the nonparametric estimation analogue. To this end, we calculate the semiparametric efficiency bound of (17) for a given weighting function $w(\cdot)$. (Note that the results with also apply for the ATE.) Since the estimate of $\Pr(T = c)$ is not affected by the weighting function, we subsequently ignore

this term. For ease of notation, we incorporate the scaling into the weighting function (13) and suppose that $c(x) = 1$. This is immaterial for the result since, for each value of x, the weighting function is anyhow re-scaled to one.

Furthermore, we note that the semiparametric efficiency bound also depends on the estimators of $\hat{\Psi}_{1i}$ and $\hat{\Psi}_{0i}$. On the other hand, for the sake of the practical feasibility of estimating $w(\cdot)$, the resulting formulae should not be too complex to prevent the need for nonparametric estimation of a large number of terms involved, which would further increase the variability of the estimated weight function. For this reason, we derive the semiparametric efficiency bound of

$$\frac{1}{n}\sum_{i=1}^{n} Y_i R_i \frac{Z_{1i} - \hat{\pi}(X_i)}{\hat{\pi}(X_i)\left(1 - \hat{\pi}(X_i)\right)} \cdot \left\{ D_i \frac{w(\Psi_{1i}, X_i)}{\Psi_{1i} \cdot \hat{f}\left(\Psi_{1i}|X_i\right)} + (1 - D_i) \frac{w(\Psi_{0i}, X_i)}{\Psi_{0i} \cdot \hat{f}\left(\Psi_{0i}|X_i\right)} \right\},$$

in which Ψ_{1i} and Ψ_{0i} are treated as covariates rather than estimated regressors. This object converges to $\tau = E\left[Y^1 - Y^0 | T = c\right] \Pr(T = c)$, which we define as

$$\tau = E\left[YR \frac{Z_1 - \pi(X)}{\pi(X)\left(1 - \pi(X)\right)} \cdot \left\{ D \frac{w(\Psi_1, X)}{\Psi_1 \cdot f\left(\Psi_1|X\right)} + (1 - D) \frac{w(\Psi_0, X)}{\Psi_0 \cdot f\left(\Psi_0|X\right)} \right\} \right].$$

In the appendix, we derive the (quite complex) influence function based on Newey (2004), which contains numerous conditional expectation terms and is therefore unlikely to be a reliable approach for estimating an appropriate weighting function in small samples (as observed in several explorative simulations). As our aim is to obtain a useful rule of thumb that works well in reasonably sized samples, we subsequently only examine the first term of the influence function, i.e. the influence function we would obtain if π and $f\left(\Psi_d|X\right)$ were known. This approach captures the direct influence of each observation on τ which does not operate via indirect estimates of nuisance parameters.

The first term of the influence function is given by

$$IF^* = YR \cdot \frac{Z_1 - \pi(X)}{\pi(X)\left(1 - \pi(X)\right)} \cdot \left\{ D \frac{w(\Psi_1, X)}{\Psi_1 \cdot f\left(\Psi_1|X\right)} + (1 - D) \frac{w(\Psi_0, X)}{\Psi_0 \cdot f\left(\Psi_0|X\right)} \right\} - \tau.$$

The semiparametric efficiency bound corresponds to the expected square of the influence function. Hence, pretending π and $f\left(\Psi_d|X\right)$ were not estimated, we obtain

$$E\left[(IF^*)^2\right] = E\left[\left(YR \cdot \frac{Z_1 - \pi(X)}{\pi(X)\left(1 - \pi(X)\right)} \cdot \left\{ D \frac{w(\Psi_1, X)}{\Psi_1 \cdot f\left(\Psi_1|X\right)} + (1 - D) \frac{w(\Psi_0, X)}{\Psi_0 \cdot f\left(\Psi_0|X\right)} \right\} - \tau \right)^2 \right],$$

which we can re-write after a few calculations (see appendix) as

$$= \int \frac{w^2(\eta, X)}{\eta^2} \left\{ \frac{\lambda_1(\eta, X)}{f_{\Psi_1|X}\left(\eta|X\right)} + \frac{\lambda_0(\eta, X)}{f_{\Psi_0|X}\left(\eta|X\right)} \right\} \cdot d\eta \cdot dF_X,$$

where

$$\lambda_1(\eta, X) = E[Y^2 RD \left(\frac{Z_1}{\pi(X)^2} + \frac{1 - Z_1}{(1 - \pi(X))^2} \right) | \Psi_1 = \eta, X], \tag{19}$$

$$\lambda_0(\eta, X) = E[Y^2 R(1 - D) \left(\frac{Z_1}{\pi(X)^2} + \frac{1 - Z_1}{(1 - \pi(X))^2} \right) | \Psi_0 = \eta, X]. \tag{20}$$

This suggests the use of the following weighting function (up to an arbitrary scaling coefficient):

Weighting function 1:

$$w(\eta, x) \propto \frac{\eta}{\sqrt{\frac{\lambda_1(\eta,x)}{f_{\Psi_1|X=x}(\eta|x)} + \frac{\lambda_0(\eta,x)}{f_{\Psi_0|X=x}(\eta|x)}}}. \tag{21}$$

As an alternative and simpler rule of thumb, we consider a function that ignores estimating the conditional means $\lambda_1(\eta, X)$ and $\lambda_0(\eta, X)$:

Weighting function 2:

$$w(\eta, x) \propto \frac{\eta}{\sqrt{\frac{1}{f_{\Psi_1|X=x}(\eta|x)} + \frac{1}{f_{\Psi_0|X=x}(\eta|x)}}}. \tag{22}$$

The latter approach only depends on estimates of $f_{\Psi_1|X}$ and $f_{\Psi_0|X}$, which need to be computed anyhow in order to inspect the common support for Ψ_1 and Ψ_0. This weighting function also has the advantage that its estimation does not make use of the data on the outcome Y, which implies that the true (and unknown) treatment effect does not affect the (estimation of the) weighting function.

4 Estimation

The identification results presented in Lemma 2 imply that the LATE may be estimated by the following expression, in which $\mathcal{S}$ denotes the set of support points of η. In principle, it may depend on x (and should then be denoted as $\mathcal{S}(x)$), but since points outside of the conditional support will anyhow receive a zero weight via the weighting function, for ease of notation we refer to $\mathcal{S}$ throughout.

$$\begin{aligned} \widehat{LATE} &= \frac{1}{\widehat{\Pr}(T = c)} \frac{1}{n} \sum_{i=1}^{n} \sum_{\eta \in \mathcal{S}} \frac{\hat{w}(\eta, X_i)}{\sum_{\eta \in \mathcal{S}} \hat{w}(\eta, X_i)} \frac{1}{\eta} \\ &\times \left(\frac{\hat{E}[YRD(Z_1 - \hat{\pi}(X))|\hat{\Psi}_1 = \eta, X_i] + \hat{E}[YR(1 - D)(Z_1 - \hat{\pi}(X))|\hat{\Psi}_0 = \eta, X_i]}{\hat{\pi}(X_i)(1 - \hat{\pi}(X_i))} \right) \end{aligned} \tag{23}$$

where

$$\widehat{\Pr}(T=c) = \frac{1}{n}\sum_{i=1}^{n} \frac{D_i}{\hat{\pi}(X_i)} \frac{Z_{1i} - \hat{\pi}(X_i)}{1-\hat{\pi}(X_i)}.$$

In the simulations and application outlined in Sections 5 and 6, the estimate $\hat{\pi}(X_i)$ of the propensity score $\Pr(Z_1 = 1|X)$ is obtained by local constant kernel regression. $\hat{\Psi}_1$ and $\hat{\Psi}_0$ are estimated based on equation (16), by a local linear regression of $R_i(Z_{1i} - \hat{\pi}(X_i)$ and $Z_{1i} - \hat{\pi}(X_i)$ on $(1, X_i, Z_{2i})$, separately among treated and non-treated observations. $\hat{E}[Y_i R_i D_i(Z_{1i} - \hat{\pi}(X_i))|\eta, X_i]$ and $\hat{E}[Y_i R_i(1 - D_i)(Z_{1i} - \hat{\pi}(X_i))|\eta, X_i]$ in (23) are obtained by a local linear regression of $Y_i R_i D_i(Z_{1i} - \hat{\pi}(X_i))$ and $Y_i R_i(1 - D_i)(Z_{1i} - \hat{\pi}(X_i))$ on $(1, \hat{\Psi}_{i1}, X_i)$ and $(1, \hat{\Psi}_{i0}, X_i)$, respectively.

Furthermore, $\hat{f}(\hat{\Psi}_1|X)$ and $\hat{f}(\hat{\Psi}_0|X)$, the conditional densities of $\hat{\Psi}_1$ and $\hat{\Psi}_0$ given X, are estimated by kernel-based density estimation, and are used as plug-in estimators for the weighting function $\hat{w}(\eta, X_i)$, which is either based on (21) or (22). For the first weighting approach, see (21), we also require estimates of λ_1 and λ_0, which we obtain by local linear regression as $Y_i^2 R_i D_i[Z_{1i}/\hat{\pi}(X_i)^2 + (1 - Z_{1i})/(1 - \hat{\pi}(X_i))^2]$ on $(1, \hat{\Psi}_{i1}, X_i)$ and $Y_i^2 R_i(1 - D_i)[Z_{1i}/\hat{\pi}(X_i)^2 + (1 - Z_{1i})/(1 - \hat{\pi}(X_i))^2]$ on $(1, \hat{\Psi}_{i0}, X_i)$ to estimate equations (19) and (20). Finally, $\mathcal{S}$ denotes the set of support points of η considered in our LATE estimator (23) which approximates the intergral over η in Lemma 2. In our simulations and applications, it consists of an equidistant 100-points grid of values starting at the maximum of (i) the minimum of $\hat{\Psi}_{i1}$, (ii) the minimum of $\hat{\Psi}_{i0}$, and (iii) 0.01 and ending at the minimum of (i) the maximum of $\hat{\Psi}_{i1}$, (ii) the maximum of $\hat{\Psi}_{i0}$, and (iii) 1.[5]

All kernel estimates (local constant/local linear regression and conditional density estimation) are based on the 'np' package of Hayfield and Racine (2008) for the statistical software R, which provides appropriate kernel functions for both continuous and discrete regressors. To be specific, we use the Gaussian kernel and the kernel function of Aitchison and Aitken (1976) for the continuous and binary regressors, respectively, in the simulations and application. The bandwidths are selected by the rule of thumb, see Silverman (1986).[6]

We also consider a semiparametric version of our estimator, in which local constant estimation is replaced by probit regression (for the propensity score) and the various local linear estimators by OLS. That is, we apply parametric first step estimators for any regression function, while the conditional densities are again estimated by (nonparametric) kernel methods.

[5]Note that in finite samples, $\hat{\Psi}_{i1}$ and $\hat{\Psi}_{i0}$ may be outside the theoretical bounds of [0,1].

[6]Using cross-validated bandwidths did not affect the results of the application much.

The ATE can be estimated in analogy to the LATE, as the former differs from the latter only in terms of weighting of the covariates:

$$\begin{aligned}\widehat{ATE} &= \frac{1}{n}\sum_{i=1}^{n}\sum_{\eta\in\mathcal{S}}\frac{\hat{w}(\eta,X_i)}{\sum_{\eta\in\mathcal{S}}\hat{w}(\eta,X_i)}\frac{1}{\eta} \\ &\times \left(\frac{\hat{E}[YRD(Z_1-\hat{\pi}(X))|\hat{\Psi}_1=\eta,X_i]+\hat{E}[YR(1-D)(Z_1-\hat{\pi}(X))|\hat{\Psi}_0=\eta,X_i]}{\hat{E}\left[DZ_1|X_i\right]-\hat{E}\left[D|X_i\right]\hat{\pi}(X_i)}\right), \quad (24)\end{aligned}$$

with $\hat{E}\left[DZ_1|X_i\right]$ and $\hat{E}\left[D|X_i\right]$ denoting estimates of $E\left[DZ_1|X_i\right]$ and $E\left[D|X_i\right]$. In addition to the regression-based estimators proposed in (23) and (24), a natural estimator using a type of IPW approach follows from equation (17):

$$\begin{aligned}\widehat{LATE} &= \frac{1}{\widehat{\Pr}(T=c)}\frac{1}{n}\sum_{i=1}^{n}Y_iR_i\frac{Z_{1i}-\hat{\pi}(X_i)}{\hat{\pi}(X_i)\left(1-\hat{\pi}(X_i)\right)\cdot c(X_i)} \\ &\times \left\{D_i\frac{w(\hat{\Psi}_{1i},X_i)}{\hat{\Psi}_{1i}\cdot\hat{f}\left(\hat{\Psi}_{1i}|X_i\right)}+(1-D_i)\frac{w(\hat{\Psi}_{0i},X_i)}{\hat{\Psi}_{0i}\cdot\hat{f}\left(\hat{\Psi}_{0i}|X_i\right)}\right\}, \quad (25)\end{aligned}$$

where $c(X_i)$ captures the scaling of the weighting function given by (13). Analogously, a natural estimator for the ATE is obtained from (18), where we also use $\Pr(T=c|X) = E\left[D\frac{Z_1-\pi(X)}{\pi(X)(1-\pi(X))}|X\right]$

$$\begin{aligned}\widehat{ATE} &= \frac{1}{n}\sum_{i=1}^{n}Y_iR_i\frac{Z_{1i}-\hat{\pi}(X_i)}{\hat{E}\left[DZ_1|X_i\right]-\hat{E}\left[D|X_i\right]\hat{\pi}(X_i)}\frac{1}{c(X_i)} \\ &\times \left\{D_i\frac{w(\hat{\Psi}_{1i},X_i)}{\hat{\Psi}_{1i}\cdot\hat{f}\left(\hat{\Psi}_{1i}|X_i\right)}+(1-D_i)\frac{w(\hat{\Psi}_{0i},X_i)}{\hat{\Psi}_{0i}\cdot\hat{f}\left(\hat{\Psi}_{0i}|X_i\right)}\right\}, \quad (26)\end{aligned}$$

with $\hat{E}\left[DZ_1|X_i\right]$ and $\hat{E}\left[D|X_i\right]$ denoting estimates of $E\left[DZ_1|X_i\right]$ and $E\left[D|X_i\right]$. In our simulations, IPW performed considerably worse than regression-based estimation such that its performance is not reported in Section 5.

5 Simulation study

To investigate the finite sample behavior of the estimator outlined in (23), we conduct a simulation study based on the following data generating process (DGP):

$$
\begin{aligned}
Y_i &= D_i - 0.5X_i + U_i, \\
Y_i &\text{ is observed if } R_i = 1, \\
R_i &= I\{D_i + \alpha X_i + Z_{2i} + V_i > 0\}, \\
D_i &= I\{Z_{1i} + \alpha X_i + W_i > 0\}, \\
Z_{1i} &= I\{\alpha X_i + P_i > 0\}, \\
Z_{2i} &= \alpha X_i + Q_i, \\
P_i, Q_i &\sim \mathcal{N}(0,1), \text{ independently of each other and of } (X_i, U_i, V_i, W_i), \\
\begin{pmatrix} U_i \\ V_i \\ W_i \end{pmatrix} &\sim \mathcal{N}(\mu, \Sigma), \text{ where } \mu = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \text{ and } \Sigma = \begin{pmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{pmatrix}
\end{aligned}
$$

The outcome variable Y_i is determined by a linear model and only observed if the binary response indicator R_i is equal to 1. The observed covariate X_i (whose distribution is specified below) is a confounder of the instruments Z_{1i} and Z_{2i}, the binary treatment D_i, the outcome, and response. Treatment endogeneity and attrition bias arise due to the nonzero covariances of U_i, V_i, W_i (given in Σ), which denote the unobserved terms in the outcome, response, and treatment equations. In contrast, P_i and Q_i, the unobservables in the instrument equations, are independent of each other and the remaining unobservables U_i, V_i, W_i, as well as X_i. Therefore, both instruments are randomly assigned given X_i.

Both (21) and (22) are considered as weighting functions in estimation based on (23). In the tables below, we refer to these as Weighting 1 (w1) and Weighting 2 (w2), respectively. As we use both nonparametric and parametric first step estimators for the various regression and density functions, this all in all entails four different estimators, denoted by 'LATE nonparametric' and 'LATE semiparametric', respectively. As a comparison, we also include the (naive) LATE estimator without attrition bias correction ('LATE naive' in the tables) based on weighting by the inverse of the nonparametric estimate of $\Pr(Z_1 = 1|X)$, see Frölich (2007). We consider two sample sizes ($n = 1000, 4000$) and two simulation designs in which X_i is either binary or continuously uniformly distributed.

Table 1: Simulations with binary covariate

	n=1000			n=4000		
	bias	st.dev.	RMSE	bias	st.dev.	RMSE
LATE nonparametric w1	-0.06	1.78	1.78	-0.01	0.52	0.52
LATE nonparametric w2	0.01	0.68	0.68	0.00	0.21	0.21
LATE semiparametric w1	-0.01	0.31	0.31	-0.01	0.15	0.15
LATE semiparametric w2	0.00	0.32	0.32	-0.00	0.16	0.16
LATE naive	-0.36	0.36	0.51	-0.35	0.17	0.39

Note: 'st.dev.' denotes the standard deviation, 'RMSE' the root mean squared error of the respective estimator. 'LATE nonparametric', 'LATE semiparametric', and 'LATE naive' refers to nonparametric estimation based on (23), semiparametric estimation based on (23) with parametric first step estimators, and LATE estimation without bias correction using IPW as outlined in Frölich (2007), respectively. 'w1' and 'w2' stands for weighting based on (21) and (22), respectively.

Table 1 presents the bias, standard deviation (st.dev.), and root mean squared error (RMSE) of the estimators for $X_i \sim \text{binom}(0.5)$. The complier and response rates are 36% and 51%, respectively, under this covariate distribution. Nonparametric estimation with non-response correction is nearly unbiased, but has a relatively large RMSE under the smaller sample size. This points to numerical instabilities of the estimator in moderate samples due to nonparametric first step estimation. Generally we find that the simpler weighting function (22), where fewer components need to be estimated, performs better. Precision and RMSE improve under the larger sample size and when using weighting function (22), the nonparametric estimator now outperforms the naive LATE, which is severely biased due to omitting non-response. However, semiparametric estimation (with non-response correction) performs considerably better w.r.t. precision and RMSE than the nonparametric method under either sample size. It even dominates naive LATE under $n = 1000$, depends little on the chosen weighting function, and appears to be the preferred choice in smaller samples.

Finally, Table 2 reports the results when the covariate follows a uniform distribution between -0.5 and 0.5 ($X_i \sim \mathcal{U}[-0.5, 0.5]$). The complier and response rates are 34% and 66%, respectively. Again, fully nonparametric estimation is relatively imprecise for $n = 1000$ (albeit less so than in the case of the binary covariate) and entails the largest RMSE. It improves as the sample size increases, but is in our DGP always outperformed by semiparametric estimation. The latter method entails

Table 2: Simulations with continuous covariate

	n=1000			n=4000		
	bias	st.dev.	RMSE	bias	st.dev.	RMSE
LATE nonparametric w1	-0.00	0.76	0.76	-0.03	0.23	0.23
LATE nonparametric w2	-0.06	0.54	0.55	-0.03	0.22	0.22
LATE semiparametric w1	0.02	0.32	0.32	0.00	0.16	0.16
LATE semiparametric w2	0.04	0.35	0.35	0.02	0.17	0.17
LATE naive	-0.33	0.32	0.46	-0.32	0.16	0.36

Note: 'st.dev.' denotes the standard deviation, 'RMSE' the root mean squared error of the respective estimator. 'LATE nonparametric', 'LATE semiparametric', and 'LATE naive' refers to nonparametric estimation based on (23), semiparametric estimation based on (23) with parametric first step estimators, and LATE estimation without bias correction using IPW as outlined in Frölich (2007), respectively. 'w1' and 'w2' stands for weighting based on (21) and (22), respectively.

RMSEs that are similar to those in Table 1 and is again stable across weighting schemes. Naive LATE estimation is once more severely biased and increasingly dominated (in terms of having a small RMSE) by the methods suggested in this paper as the sample size grows. For the sample sizes examined here semiparametric estimation works best.

6 Application: The effects of sports on self-reported health

6.1 The experiment

The estimator outlined in (23) is applied in a field experiment to analyze the short-term effect of recreational sport and exercise in university on self-assessed health. Campus sports and exercise are an integral part of university life. Universities usually offer these programs and facilities to promote a healthy and balanced lifestyle of their students. While in general health benefits of sports and physical exercise are well established,[7] little is known about the health effects of recreational campus sports and exercise. A fundamental problem of this literature is the self-selection into sports. Students who practice sports potentially differ in observable and unobservable char-

[7]See Timmons, Leblanc, Carson, Connor Gorber, Dillman, Janssen, Kho, Spence, Stearns, and Tremblay (2012) for small children, Janssen and Leblanc (2010) for adolescents, and Reiner, Niermann, Jekauc, and Woll (2013) for adults.

acteristics from those students that do not.[8] To solve this endogeneity problem, Fricke, Lechner, and Steinmayr (2015) carried out an experiment at the University of St.Gallen,[9] in which they randomly assigned incentives to exercise among students. Specifically, they provided first year students in the cohort 2013 who participated in a baseline survey (sample size $n = 472$) randomly with cash incentives to participate in campus sports and exercise.[10] Half of the students received a cash incentive of 100 CHF, while the other half did not. (At the time of the experiment, this amount was approximately worth USD 110.)[11]

Using this experiment, we randomly invited students to enter a lottery in order to incentivize participation in a follow-up survey which measures self-reported health.[12] Students could win a cash price with a chance of 25% conditional on survey participation. The cash prices varied in steps of CHF 10 from 10 to CHF 200 and each of them was offered to approximately 20 students. The survey was sent to the students at the end of the second semester. Additionally, students received up to four reminders to participate in the survey. We sent an email offering a lottery conditional on survey participation to some students after the first survey email and after the fourth reminder.[13] Note that the lottery was randomized among students who were still enrolled at the university.

The research design makes use of three different data sources. First, the treatment is based on data from the university ID scanner at the entrance of the university gym. This gym covers most of the university's sports and exercise activities.[14] Second, the administrative student records of the university provide us with socio-demographic information such as gender, age, nationality, and mother tongue. Third, the outcome, self-reported health which ranges from (1) very good to (5)

[8] See for example Schneider and Becker (2005), and Farrell and Shields (2002).

[9] The University of St.Gallen is one of 12 public universities in Switzerland. Its covers the fields of Business Administration, Economics, International Affairs, and Law. In 2013, it accommodates approximately 7700 students.

[10] Charness and Gneezy (2009) document the the effectiveness of cash incentives to increase physical activity.

[11] The exact implementation was as follows: The students were split into 13 blocks conditional on individual characteristics. In all blocks, approximately half of the students were assigned to the treatment group and to the control group. If students use the campus sports and exercise facilities twice per week over ten weeks, they receive the entire amount. Each week the endowment is reduced by CHF 5 if they participate only once a week, or by CHF 10 if they do not participate at all.

[12] Again, we randomized within three blocks conditional on individual characteristics.

[13] The lottery email was sent directly after the survey email in order avoid an additional reminder effect of the lottery.

[14] Activities include besides the regular gym, a multitude of courses, as well as team sports.

poor, is taken from the follow-up survey at the end of the second semester.[15]

6.2 Descriptive statistics

Table 3 shows descriptive statistics in the total sample as well as conditional on (not) receiving a cash incentive and lottery offer. The sample consists of mostly Swiss (81%), German speaking (90%) students. Thirty-seven percent of the students are female and the average age at enrollment is approximately 20 years. Moreover, Table 3 allows assessing the quality of the randomization of both instruments. Column (4) provides the mean differences of student characteristics, the health outcome, the treatment 'one or more gym visits', and follow-up response across the groups with and without cash incentive. The respective p-values suggest that the student characteristics are well balanced. Column (7) gives the mean differences of the previous variables as well as the cash incentive instrument across students receiving no lottery offer and some offer larger than zero. Column (8) provides an F-test for joint significance of the amount of the cash lottery and its square in an OLS or probit regression of the respective non-binary or binary variable. The results suggest that both the student characteristics and the cash incentives are comparable across cash lottery recipients and non-recipients.

We now consider the effectiveness of either instrument. The probability of students who receive the cash incentives to visit the gym at least once is 82.2%, which 7.9 percentage points higher than among students not receiving the incentives (74.3%). The difference is significant at the 5% level (using heteroskedasticity robust standard errors). As for the cash lottery for response, the offer of a positive value increases the follow-up survey response rate by 23 percentage points, i.e. from 48% to 72%. Furthermore, Figure 1 suggests that the response rate increases nonlinearly with the value of the lottery, with the strongest marginal effects between CHF 80 and 140. The response rates reach around 80% for high lottery values of CHF 140 to 200.

6.3 Results

The (binary) treatment is defined as visiting the university gym at least once during the first study year. In our estimations, we condition on the (binary) covariates gender ('female') and Swiss nationality ('Swiss'). The first stage effect is 0.084, implying that the cash incentive instrument

[15]For only two out of all respondents of the follow-up survey, self-reported health is missing (item non-response). The two missing values were set to the mean of self-assessed health among the 470 survey respondents without item non-response. Alternatively, deleting these two observations led to qualitatively similar conclusions.

Table 3: Descriptive Statistics

		Cash Incentive to visit gym			Lottery incentive for response			
	Overall	Control	Cash incentives	Difference (p-value)	Lottery = 0	Lottery > 0	Difference (p-value)	F-Statistic (p-value)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Swiss	0.81	0.81	0.81	-0.01	0.80	0.81	0.01	0.32
				(0.84)			(0.88)	(0.73)
German mother tongue	0.90	0.92	0.89	-0.03	0.92	0.90	-0.02	0.28
				(0.29)			(0.58)	(0.75)
Female	0.37	0.37	0.37	0.00	0.39	0.36	-0.03	0.66
				(0.97)			(0.6)	(0.52)
Age	19.87	19.88	19.85	-0.03	19.86	19.87	0.01	0.20
				(0.81)			(0.93)	(0.82)
Self-reported health	1.92	1.90	1.95	0.05	1.93	1.92	-0.00	0.12
				(0.38)			(0.95)	(0.89)
Response incentives	-	-	-	-	0.46	0.53	0.06	0.17
				-			(0.28)	(0.84)
One or more gym visits	0.78	0.74	0.82	0.08	0.72	0.80	0.08	0.34
				(0.04)			(0.12)	(0.71)
Follow-up response	0.67	0.66	0.68	0.02	0.48	0.72	0.23	13.95
				(0.70)			(0.00)	(0.00)
N		230	242		97	375		

Note: Column (1) shows the overall sample mean. Columns (2), (3), (5), and (6) give the respective group means. The F-statistic corresponds to an F-test for joint significance of the amount of the cash lottery and its square in an OLS or probit regression of the respective non-binary or binary attribute. 'Swiss' is a binary indicator for Swiss nationality. 'German mother tongue' is a binary indicator for native language of the student. 'Age' refers to the age at enrollment. 'Self-reported health' ranges from 1: very good to 5: poor. The corresponding statistics are only reported for students who answer the follow up survey. 'Cash incentives to visit gym' is a binary indicator for the receipt of the incentives to exercise. 'One or more gym visits' is a binary indicator for visiting the gym at least once over the two semesters. 'Follow-up response' is a binary indicator for participation in the follow-up survey.

increases the probability to visit the gym at least once by 8.4 percentage points (which corresponds to the complier share),[16] with a bootstrap p-value of 0.056. Table 4 reports the LATE estimates based on (23) using the semiparametric approach, which performed much better in the previous

[16]This number differs slightly from that in Table 3 since we control for covariates X.

Figure 1: Follow up Survey Response

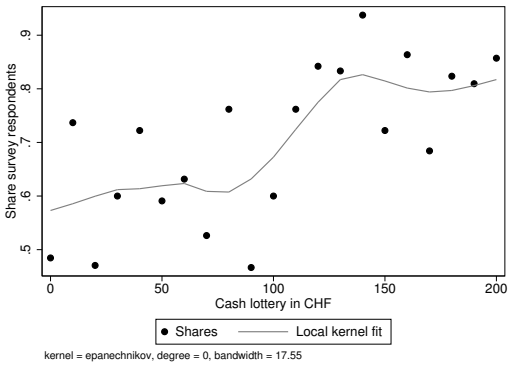

Table 4: Application

effects	estimate	p-value	differences in effects	estimate	p-value
LATE semiparametric w1	0.91	0.87	semiparametric w1 − naive	0.08	0.93
LATE semiparametric w2	1.00	0.91	semiparametric w2 − naive	0.18	0.84
LATE naive (no attrition correction)	0.83	0.47			

Note: P-values are based on the quantiles of the bootstrapped effects using 1999 bootstrap replications. 'LATE semiparametric' and 'LATE naive' refers to semiparametric estimation based on (23) with parametric first step estimators and LATE estimation without bias correction using IPW as outlined in Frölich (2007), respectively. 'w1' and 'w2', stands for weighting based on (21) and (22), respectively. The outcome 'self-reported health' ranges from 1: very good to 5: poor.

Monte Carlo simulation than fully nonparametric estimation (which appeared to be very noisy in small samples). The table reports 'LATE semiparametric', both with weighting functions (21) and (22), see 'w1' and 'w2'. Also included is the nonparametric LATE estimator without attrition correction based on IPW ('LATE naive') as outlined in Frölich (2007), as well as the differences between the latter and the former approaches ('semiparametric w1 − naive ...'). This estimator is inconsistent unless attrition was a non-selective random event, i.e. only depending on observables. Besides the LATEs, the table gives the bootstrap p-values based on the quantiles of the resampled distribution of the effect estimates (1999 replications), see eq. (6) in MacKinnon (2006). We provide the quantile-based p-values (rather than those based on the t-statistic) to account for the problem that in finite samples the moments of instrumental variable estimators may not exist such that t-statistics may be misleading.

The naive LATE estimator is positive (0.83 points), implying a decrease in self-assessed health if taken at face value, but at the same time far from being significant at any conventional level. With point estimates 0.91 and 1.00, the attrition-corrected estimates are similar and insignificant, too. So are the differences between the bias corrected and naive LATE estimators. We conclude that in our application, we find neither evidence for an effect of gym training on (short term) self-assessed health nor for selective attrition (with respect to the unobserved potential outcomes). We acknowledge that this may be due to the low precision of our estimates rooted in the small sample size and complier share.

7 Conclusion

This paper developed a nonparametric identification approach of average treatment effects in the presence of both treatment endogeneity and attrition/non-response, using a discrete instrument for the binary treatment and a continuous instrument for attrition. Furthermore, we proposed nonparametric and semiparametric estimators based on the sample analogs of our identification results and investigated their performance in a small simulation study. As an empirical illustration, we considered a randomized experiment at a Swiss University in order to estimate the effect of gym training on students' self-assessed health, where the treatment (gym training) and attrition were instrumented by randomized cash incentives (paid out conditional on gym visits) and by a cash lottery for participating in the follow-up survey, respectively. We did not find evidence for selective attrition. In future research we will also examine settings for non-binary treatments, which presumably will require more demanding identification approaches in requiring several continuous instrumental variables (which may be harder to find in applied research).

References

Abadie, A. (2003): "Semiparametric instrumental Variable estimation of treatment response models," *Journal of Econometrics*, 113, 231–263.

Abowd, J., B. Crepon, and F. Kramarz (2001): "Moment Estimation With Attrition: An Application to Economic Models," *Journal of the American Statistical Association*, 96, 1223–1230.

Aitchison, J., and C. Aitken (1976): "Multivariate binary discrimination by the kernel method," *Biometrika*, 63, 413–420.

Angrist, J., and I. Fernández-Val (2010): "Extrapolate-ing: External validity and overidentification in the late framework," *NBER working paper 16566*.

Angrist, J., G. Imbens, and D. Rubin (1996): "Identification of Causal Effects using Instrumental Variables," *Journal of American Statistical Association*, 91, 444–472 (with discussion).

Behaghel, L., B. Crépon, M. Gurgand, and T. Le Barbanchon (2012): "Please call again: Correcting non-response bias in treatment effect models," *IZA Discussion Paper No. 6751*.

Castiglioni, L., K. Pforr, and U. Krieger (2008): "The Effect of Incentives on Response Rates and Panel Attrition: Results of a Controlled Experiment," *Survey Research Methods*, 2, 151–158.

Charness, G., and U. Gneezy (2009): "Incentives to Exercise," *Econometrica*, 77(3), 909–931.

DiNardo, J., J. McCrary, and L. Sanbonmatsu (2006): "Constructive Proposals for Dealing with Attrition: An Empirical Example," *Working paper, University of Michigan*.

Farrell, L., and M. A. Shields (2002): "Investigating the economic and demographic determinants of sporting participation in England," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 165(2), 335–348.

Fitzgerald, J., P. Gottschalk, and R. Moffitt (1998): "An Analysis of Sample Attrition in Panel Data: The Michigan Panel Study of Income Dynamics," *Journal of Human Resources*, 33, 251–299.

Frangakis, C., and D. Rubin (1999): "Addressing complications of intention-to-treat analysis in the combined presence of all-or-none treatment-noncompliance and subsequent missing outcomes," *Biometrika*, 86, 365–379.

Fricke, H., M. Lechner, and A. Steinmayr (2015): "Effects of university sports and excercise on health and eductional outcomes: Evidence from a randomized experiment," *mimeo, University of St. Gallen*.

Frölich, M. (2007): "Nonparametric IV Estimation of Local Average Treatment Effects with Covariates," *Journal of Econometrics*, 139, 35–75.

Frölich, M., and M. Huber (2014): "Treatment evaluation with multiple outcome periods under endogeneity and attrition," *Journal of the American Statistical Association*, 109, 1697–1711.

Frumento, P., F. Mealli, B. Pacini, and D. B. Rubin (2012): "Evaluating the Effect of Training on Wages in the Presence of Noncompliance, Nonemployment, and Missing Outcome Data," *Journal of the American Statistical Association*, 107, 450–466.

Hausman, J., and D. Wise (1979): "Attrition Bias In Experimental and Panel Data: The Gary Income Maintenance Experiment," *Econometrica*, 47(2), 455–473.

Hayfield, T., and J. Racine (2008): "Nonparametric Econometrics: The np Package," *Journal of Statistical Software*, 27, 1–32.

Heckman, J. (1976): "The Common Structure of Statistical Models of Truncation, Sample Selection, and Limited Dependent Variables, and a Simple Estimator for such Models," *Annals of Economic and Social Measurement*, 5, 475–492.

Heckman, J. (1979): "Sample Selection Bias as a Specification Error," *Econometrica*, 47, 153–161.

Horvitz, D. G., and D. J. Thompson (1952): "A Generalization of Sampling without Replacement from a Finite Universe," *Journal of the American Statistical Association*, 47, 663–685.

Huber, M. (2012): "Identification of average treatment effects in social experiments under alternative forms of attrition," *Journal of Educational and Behavioral Statistics*, 37, 443–474.

Huber, M. (2014): "Treatment evaluation in the presence of sample selection," *Econometric Reviews*, 33, 869–905.

Huber, M., and G. Mellace (2015): "Testing Instrument Validity for LATE Identification Based on Inequality Moment Constraints," *Review of Economics and Statistics*, 97, 398–411.

Imbens, G., and J. Angrist (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62, 467–475.

Janssen, I., and A. G. Leblanc (2010): "Systematic review of the health benefits of physical activity and fitness in school-aged children and youth," *The International Journal of Behavioral Nutrition and Physical Activity*, 7, 40.

Kitagawa, T. (2015): "A Test for Instrument Validity," *Econometrica*, 83, 2043–2063.

Little, R., and D. Rubin (1987): *Statistical Analysis with Missing Data*. Wiley, New York.

MacKinnon, J. G. (2006): "Bootstrap Methods in Econometrics," *The Economic Record*, 82, S2–S18.

Mealli, F., G. Imbens, S. Ferro, and A. Biggeri (2004): "Analyzing a randomized trial on breast self-examination with noncompliance and missing outcomes," *Biostatistics*, 5, 207–222.

Newey, W. (2004): "Efficient semiparametric estimation via moment restrictions," *Econometrica*, 72, 1877–1897.

Reiner, M., C. Niermann, D. Jekauc, and A. Woll (2013): "Long-term health benefits of physical activity–a systematic review of longitudinal studies," *BMC public health*, 13, 813.

Robins, J. M., A. Rotnitzky, and L. Zhao (1994): "Estimation of Regression Coefficients When Some Regressors Are not Always Observed," *Journal of the American Statistical Association*, 90, 846–866.

Rubin, D. (1974): "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, 66, 688–701.

——— (1976): "Inference and Missing Data," *Biometrika*, 63, 581–592.

Schneider, S., and S. Becker (2005): "Prevalence of Physical Activity among the Working Population and Correlation with Work-Related Factors: Results from the First German National Health Survey," *Journal of Occupational Health*, 47(5), 414–423.

Schwiebert, J. (2012): "Semiparametric Estimation of a Sample Selection Model in the Presence of Endogeneity," *unpublished manuscript*.

Semykina, A., and J. Wooldridge (2010): "Estimating Panel Data Models in the Presence of Endogeneity and Selection: Theory and Application," *Journal of Econometrics*, 157, 375–380.

Silverman, B. (1986): *Density estimation for statistics and data analysis*. Chapman and Hall, London.

Tan, Z. (2006): "Regression and Weighting Methods for Causal Inference Using Instrumental Variables," *Journal of the American Statistical Association*, 101, 1607–1618.

Timmons, B. W., A. G. Leblanc, V. Carson, S. Connor Gorber, C. Dillman, I. Janssen, M. E. Kho, J. C. Spence, J. A. Stearns, and M. S. Tremblay (2012): "Systematic review of physical activity and health in the early years (aged 0-4 years)," *Applied Physiology, Nutrition, and Metabolism = Physiologie Appliquée, Nutrition Et Métabolisme*, 37(4), 773–792.

Zhang, J., D. Rubin, and F. Mealli (2009): "Likelihood-Based Analysis of Causal Effects of Job-Training Programs Using Principal Stratification," *Journal of the American Statistical Association*, 104, 166–176.

A Appendix: Proofs of Theorems

A.1 Preliminaries

We will repeatedly make use of

$$\begin{aligned} & E\left[\frac{D}{p(Z_2,X)}\frac{Z_1-p(Z_2,X)}{1-p(Z_2,X)}|X,Z_2\right] \\ = \;& E\left[D|X,Z_2,Z_1=1\right]-E\left[D|X,Z_2,Z_1=0\right] \\ = \;& \Pr\left(T=c|X,Z_2\right). \end{aligned} \tag{A.1}$$

The proof is immediate via partitioning by types, as

$$\begin{aligned} & E\left[D|X,Z_2,Z_1=1\right] \\ = \;& E\left[D|X,Z_2,Z_1=1,T=a\right]\Pr\left(T=a|X,Z_2,Z_1=1\right) \\ & +E\left[D|X,Z_2,Z_1=1,T=c\right]\Pr\left(T=c|X,Z_2,Z_1=1\right) \\ & +E\left[D|X,Z_2,Z_1=1,T=n\right]\Pr\left(T=n|X,Z_2,Z_1=1\right) \\ = \;& \Pr\left(T=a|X,Z_2,Z_1=1\right)+\Pr\left(T=c|X,Z_2,Z_1=1\right) \\ = \;& \Pr\left(T=a|X,Z_2\right)+\Pr\left(T=c|X,Z_2\right), \end{aligned}$$

because of $T \perp\!\!\!\perp Z_1|X,Z_2$. With analogous derivations for $E\left[D|X,Z_2,Z_1=0\right]$, the result follows immediately.

A.2 Proof of Lemma 1

We show the result for $\psi_1(z_2, x)$ and note that the derivations for $\psi_0(z_2, x)$ are analogous. Note that

$$\begin{aligned}
& E\left[RD|X=x, Z_2=z_2, Z_1=1\right] \\
= \; & E\left[RD|X=x, Z_2=z_2, Z_1=1, T=c\right] \Pr\left(T=c|X=x, Z_2=z_2, Z_1=1\right) \\
& + E\left[RD|X=x, Z_2=z_2, Z_1=1, T=a\right] \Pr\left(T=a|X=x, Z_2=z_2, Z_1=1\right) \\
& + E\left[RD|X=x, Z_2=z_2, Z_1=1, T=n\right] \Pr\left(T=n|X=x, Z_2=z_2, Z_1=1\right). \\
= \; & \Pr\left(\zeta(1, z_2, x) \geq V \; |X=x, Z_2=z_2, Z_1=1, T=c\right) \Pr\left(T=c|X=x, Z_2=z_2\right) \\
& + \Pr\left(\zeta(1, z_2, x) \geq V \; |X=x, Z_2=z_2, Z_1=1, T=a\right) \Pr\left(T=a|X=x, Z_2=z_2\right) \\
= \; & F_{V|X=x, Z_2=z_2, Z_1=1, T=c}\left(\zeta(1, z_2, x)\right) \Pr\left(T=c|X=x, Z_2=z_2\right) \\
& + F_{V|X=x, Z_2=z_2, Z_1=1, T=a}\left(\zeta(1, z_2, x)\right) \Pr\left(T=a|X=x, Z_2=z_2\right) \\
= \; & F_{V|X=x, T=c}\left(\zeta(1, z_2, x)\right) \cdot \Pr\left(T=c|X=x, Z_2=z_2\right) \\
& + F_{V|X=x, T=a}\left(\zeta(1, z_2, x)\right) \cdot \Pr\left(T=a|X=x, Z_2=z_2\right),
\end{aligned}$$

where the second equality follows from inserting the definition of the types and using $T \perp\!\!\!\perp Z_1 | X, Z_2$ and the fourth equality follows from $V \perp\!\!\!\perp (Z_1, Z_2) | X, T$

With this intermediary result and with analogous derivations for $E\left[RD|X, Z_2, Z_1=0\right]$ we obtain

$$\begin{aligned}
& E\left[RD|X=x, Z_2=z_2, Z_1=1\right] - E\left[RD|X=x, Z_2=z_2, Z_1=0\right] \\
= \; & F_{V|X=x, T=c}\left(\zeta(1, z_2, x)\right) \cdot \Pr\left(T=c|X=x, Z_2=z_2\right)
\end{aligned}$$

$$= E\left[\frac{RD}{E\left[Z_1|X=x, Z_2=z_2\right]} \frac{Z_1 - E\left[Z_1|X=x, Z_2=z_2\right]}{1 - E\left[Z_1|X=x, Z_2=z_2\right]} | X=x, Z_2=z_2\right].$$

Now inserting the result of (A.1) for $\Pr\left(T=c|X=x, Z_2=z_2\right)$ we obtain

$$\frac{E\left[\frac{RD}{p(z_2,x)} \frac{Z_1 - p(z_2,x)}{1-p(z_2,x)} | X=x, Z_2=z_2\right]}{E\left[\frac{D}{p(z_2,x)} \frac{Z_1 - p(z_2,x)}{1-p(z_2,x)} | X=x, Z_2=z_2\right]} = F_{V|X=x, T=c}\left(\zeta(1, z_2, x)\right),$$

which simplifies to

$$\begin{aligned}
= \; & \frac{E\left[RD \cdot \left(Z_1 - p\left(z_2, x\right)\right) | X=x, Z_2=z_2\right]}{E\left[D \cdot \left(Z_1 - p\left(z_2, x\right)\right) | X=x, Z_2=z_2\right]} \\
= \; & \frac{E\left[R \cdot \left(Z_1 - p\left(z_2, x\right)\right) | D=1, X=x, Z_2=z_2\right]}{E\left[Z_1 - p\left(z_2, x\right) | D=1, X=x, Z_2=z_2\right]}.
\end{aligned}$$

A.3 Proof of Theorem 1

Consider some value x and suppose that $\mathcal{X}_x$ is non-empty. Let $\bar{\eta} \in \mathcal{X}_x$ be a value from the common support. Now consider the expression

$$E\left[YRD \frac{\Pr\left(Z_1=1|X=x, \Psi_1=\bar{\eta}\right)}{\Pr\left(Z_1=1|Z_2, X=x, \Psi_1=\bar{\eta}\right)} | X=x, \Psi_1=\bar{\eta}, Z_1=1\right] \tag{A.2}$$

$$
\begin{aligned}
&= \int E\left[YRD\frac{\Pr\left(Z_1=1|X=x,\Psi_1=\bar{\eta}\right)}{\Pr\left(Z_1=1|Z_2,X=x,\Psi_1=\bar{\eta}\right)}|Z_2,X=x,\Psi_1=\bar{\eta},Z_1=1\right]dF_{Z_2|X=x,\Psi_1=\bar{\eta},Z_1=1}\\
&= \int E\left[YRD|Z_2,X=x,\Psi_1=\bar{\eta},Z_1=1\right]dF_{Z_2|X=x,\Psi_1=\bar{\eta}}\\
&= \int E\left[YRD|Z_2,X=x,Z_1=1\right]dF_{Z_2|X=x,\Psi_1=\bar{\eta}},
\end{aligned}
$$

which follows from Bayes theorem and because $\Psi_1=\psi_1(Z_2,X)$ is a function of Z_2 and X only. Now partitioning by type, inserting the model, and using $T\perp\!\!\!\perp Z_1|X,Z_2$ and $(U,V)\perp\!\!\!\perp(Z_1,Z_2)|X,T$ implies:

$$
\begin{aligned}
&\int E\left[YRD|Z_2,X=x,Z_1=1\right]dF_{Z_2|X=x,\Psi_1=\bar{\eta}}\\
=&\int E\left[\{\varphi(1,x)+U\}\cdot 1\left(\zeta(1,z_2,x)\geq V\right)|Z_2=z_2,X=x,Z_1=1,T=a\right]\Pr\left(T=a|Z_2=z_2,X=x\right)dF_{Z_2|X=x,\Psi_1=\bar{\eta}}\\
+&\int E\left[\{\varphi(1,x)+U\}\cdot 1\left(\zeta(1,z_2,x)\geq V\right)|Z_2=z_2,X=x,Z_1=1,T=c\right]\Pr\left(T=c|Z_2=z_2,X=x\right)dF_{Z_2|X=x,\Psi_1=\bar{\eta}}\\
=&\int E\left[\{\varphi(1,x)+U\}\cdot 1\left(\zeta(1,z_2,x)\geq V\right)|X=x,T=a\right]\Pr\left(T=a|Z_2=z_2,X=x\right)dF_{Z_2|X=x,\Psi_1=\bar{\eta}}\\
+&\int E\left[\{\varphi(1,x)+U\}\cdot 1\left(\zeta(1,z_2,x)\geq V\right)|X=x,T=c\right]\Pr\left(T=c|Z_2=z_2,X=x\right)dF_{Z_2|X=x,\Psi_1=\bar{\eta}}.
\end{aligned}
$$

Analogously, we can derive a similar expression as (A.2) for the $Z_1=0$ subpopulation. Combining the two results we obtain:

$$
\begin{aligned}
&E\left[YRD\frac{\Pr\left(Z_1=1|X=x,\Psi_1=\bar{\eta}\right)}{\Pr\left(Z_1=1|Z_2,X=x,\Psi_1=\bar{\eta}\right)}|X=x,\Psi_1=\bar{\eta},Z_1=1\right]\\
&\qquad - E\left[YRD\frac{\Pr\left(Z_1=0|X=x,\Psi_1=\bar{\eta}\right)}{\Pr\left(Z_1=0|Z_2,X=x,\Psi_1=\bar{\eta}\right)}|X=x,\Psi_1=\bar{\eta},Z_1=0\right] \qquad \text{(A.3)}
\end{aligned}
$$

$$
\begin{aligned}
&= \int E\left[\{\varphi(1,x)+U\}\cdot 1\left(\zeta(1,z_2,x)\geq V\right)|X=x,T=c\right]\Pr\left(T=c|Z_2=z_2,X=x\right)dF_{Z_2|X=x,\Psi_1=\bar{\eta}}\\
&= \int E\left[\{\varphi(1,x)+U\}\cdot 1\left(V\leq F^{-1}_{V|X=x,T=c}(\bar{\eta})\right)|X=x,T=c\right]\Pr\left(T=c|Z_2=z_2,X=x\right)dF_{Z_2|X=x,\Psi_1=\bar{\eta}}
\end{aligned}
$$

which follows from Lemma 1, with $F^{-1}_{V|X=x,T=c}$ being the inverse function of $F_{V|X=x,T=c}$. Again using that $\Psi_1=\psi_1(Z_2,X)$ is a function of Z_2 and X only we can also see that the last terms in the previous expression simplify to $\Pr\left(T=c|X=x,\Psi_1=\bar{\eta}\right)$, such that we obtain

$$
\begin{aligned}
&E\left[\{\varphi(1,x)+U\}\cdot 1\left(V\leq F^{-1}_{V|X=x,T=c}(\bar{\eta})\right)|X=x,T=c\right]\Pr\left(T=c|X=x,\Psi_1=\bar{\eta}\right)\\
=&\left\{\varphi(1,x)\cdot\bar{\eta}+\int E\left[U\cdot 1\left(V\leq F^{-1}_{V|X=x,T=c}(\bar{\eta})\right)|X=x,T=c\right]\right\}\Pr\left(T=c|X=x,\Psi_1=\bar{\eta}\right) \qquad \text{(A.4)}
\end{aligned}
$$

Note that we can also re-write expression (A.3) as

$$
\begin{aligned}
&= E\left[YRD\frac{Z_1}{\Pr\left(Z_1=1|Z_2,X=x,\Psi_1=\bar{\eta}\right)}|X=x,\Psi_1=\bar{\eta}\right]\\
&\qquad - E\left[YRD\frac{1-Z_1}{\Pr\left(Z_1=0|Z_2,X=x,\Psi_1=\bar{\eta}\right)}|X=x,\Psi_1=\bar{\eta}\right]
\end{aligned}
$$

$$= E\left[\frac{YRD}{E\left[Z_1|Z_2, X=x, \Psi_1=\bar{\eta}\right]}\frac{Z_1 - E\left[Z_1|Z_2, X=x, \Psi_1=\bar{\eta}\right]}{1-E\left[Z_1|Z_2, X=x, \Psi_1=\bar{\eta}\right]}|X=x, \Psi_1=\bar{\eta}\right]. \tag{A.5}$$

By analogous derivations,

$$E\left[\frac{D}{E\left[Z_1|Z_2, X=x, \Psi_1=\bar{\eta}\right]}\frac{Z_1 - E\left[Z_1|Z_2, X=x, \Psi_1=\bar{\eta}\right]}{1-E\left[Z_1|Z_2, X=x, \Psi_1=\bar{\eta}\right]}|X=x, \Psi_1=\bar{\eta}\right] = \Pr\left(T=c|X=x, \Psi_1=\bar{\eta}\right). \tag{A.6}$$

Analogously, we can derive

$$E\left[\frac{YR(1-D)}{E\left[Z_1|Z_2, X, \Psi_0=\bar{\eta}\right]}\frac{Z_1 - E\left[Z_1|Z_2, X, \Psi_0=\bar{\eta}\right]}{1-E\left[Z_1|Z_2, X, \Psi_0=\bar{\eta}\right]}|X=x, \Psi_0=\bar{\eta}\right] \tag{A.7}$$

$$\begin{aligned}= E\left[YR(1-D)\frac{\Pr\left(Z_1=1|X=x, \Psi_0=\bar{\eta}\right)}{\Pr\left(Z_1=1|Z_2, X=x, \Psi_0=\bar{\eta}\right)}|X=x, \Psi_0=\bar{\eta}, Z_1=1\right] \\ - E\left[YR(1-D)\frac{\Pr\left(Z_1=0|X=x, \Psi_0=\bar{\eta}\right)}{\Pr\left(Z_1=0|Z_2, X=x, \Psi_0=\bar{\eta}\right)}|X=x, \Psi_0=\bar{\eta}, Z_1=0\right]\end{aligned}$$

$$= -\left\{\varphi(0,x)\cdot\bar{\eta} + \int E\left[U\cdot 1\left(V \leq F^{-1}_{V|X=x,T=c}(\bar{\eta})\right)|X=x, T=c\right]\right\}\Pr\left(T=c|X=x, \Psi_0=\bar{\eta}\right). \tag{A.8}$$

Similarly, we can derive

$$E\left[\frac{1-D}{E\left[Z_1|Z_2, X, \Psi_0=\bar{\eta}\right]}\frac{Z_1 - E\left[Z_1|Z_2, X, \Psi_0=\bar{\eta}\right]}{1-E\left[Z_1|Z_2, X, \Psi_0=\bar{\eta}\right]}|X=x, \Psi_0=\bar{\eta}\right] = -\Pr\left(T=c|X=x, \Psi_0=\bar{\eta}\right). \tag{A.9}$$

Now putting all results together, we obtain

$$\begin{aligned}&\frac{1}{\bar{\eta}}\frac{E\left[\frac{YRD}{E[Z_1|Z_2,X=x,\Psi_1=\bar{\eta}]}\frac{Z_1-E[Z_1|Z_2,X=x,\Psi_1=\bar{\eta}]}{1-E[Z_1|Z_2,X=x,\Psi_1=\bar{\eta}]}|X=x,\Psi_1=\bar{\eta}\right]}{E\left[\frac{D}{E[Z_1|Z_2,X=x,\Psi_1=\bar{\eta}]}\frac{Z_1-E[Z_1|Z_2,X=x,\Psi_1=\bar{\eta}]}{1-E[Z_1|Z_2,X=x,\Psi_1=\bar{\eta}]}|X=x,\Psi_1=\bar{\eta}\right]} \\ &\qquad - \frac{1}{\bar{\eta}}\frac{E\left[\frac{YR(1-D)}{E[Z_1|Z_2,X=x,\Psi_0=\bar{\eta}]}\frac{Z_1-E[Z_1|Z_2,X=x,\Psi_0=\bar{\eta}]}{1-E[Z_1|Z_2,X=x,\Psi_0=\bar{\eta}]}|X=x,\Psi_0=\bar{\eta}\right]}{E\left[\frac{1-D}{E[Z_1|Z_2,X=x,\Psi_0=\bar{\eta}]}\frac{Z_1-E[Z_1|Z_2,X=x,\Psi_0=\bar{\eta}]}{1-E[Z_1|Z_2,X=x,\Psi_0=\bar{\eta}]}|X=x,\Psi_0=\bar{\eta}\right]}\end{aligned}$$

$$\begin{aligned}= \quad & \frac{1}{\bar{\eta}}\left\{\varphi(1,x)\cdot\bar{\eta} + \int E\left[U\cdot 1\left(V \leq F^{-1}_{V|X=x,T=c}(\bar{\eta})\right)|X=x, T=c\right]\right\} \\ & -\frac{1}{\bar{\eta}}\left\{\varphi(0,x)\cdot\bar{\eta} + \int E\left[U\cdot 1\left(V \leq F^{-1}_{V|X=x,T=c}(\bar{\eta})\right)|X=x, T=c\right]\right\} \\ = \varphi(1,x) - \varphi(0,x) &= E\left[Y^1 - Y^0|X=x, T=c\right] = E\left[Y^1 - Y^0|X=x\right].\end{aligned}$$

Define

$$\Xi_d(x,\eta) = \frac{E\left[\frac{YR}{E[Z_1|Z_2,X=x,\Psi_d=\eta]}\frac{Z_1-E[Z_1|Z_2,X=x,\Psi_d=\eta]}{1-E[Z_1|Z_2,X=x,\Psi_d=\eta]}|D=d,X=x,\Psi_d=\eta\right]}{E\left[\frac{1}{E[Z_1|Z_2,X=x,\Psi_d=\eta]}\frac{Z_1-E[Z_1|Z_2,X=x,\Psi_d=\eta]}{1-E[Z_1|Z_2,X=x,\Psi_d=\eta]}|D=d,X=x,\Psi_d=\eta\right]}.$$

Now we obtain

$$\frac{1}{\eta}\left(\ \Xi_1(x,\eta)-\Xi_0(x,\eta)\ \right)=E\left[Y^1-Y^0|X=x\right]=E\left[Y^1-Y^0|X=x,T=c\right]$$

.

In principle, a single value $\eta\in\mathcal{X}_x$ suffices for identification of the treatment effect conditional on X. For estimation, this would, however, imply that only a rather limited amount of the information in the data was used. Instead, we might consider all values $\eta\in\mathcal{X}_x\subseteq[0,1]$ and choose some weighting scheme $w(\eta,x)$ as a function of η and possibly also of x. One may therefore identify the conditional treatment effect as

$$=\frac{\int_0^1\frac{1}{\eta}\left(\ \Xi_1(x,\eta)-\Xi_0(x,\eta)\ \right)w(\eta,x)d\eta}{\int_0^1 w(\eta,x)d\eta},$$

provided that the weighting function $w(\eta,x)$ does not integrate to zero. With this result we obtain the average treatment effect as

$$E\left[Y^1-Y^0\right]=\int\left(\int_0^1\frac{1}{\eta}\left(\Xi_1(X,\eta)-\Xi_0(X,\eta)\ \right)\frac{w(\eta,X)}{\int w(\eta,X)d\eta}\right)d\eta dF_X$$

and the local average treatment effect on the compliers as

$$\begin{aligned}E\left[Y^1-Y^0|T=c\right] &= \int E\left[Y^1-Y^0|X,T=c\right]dF_{X|T=c}\\ &= \int E\left[Y^1-Y^0|X,T=c\right]\frac{\Pr\left(T=c|X\right)dF_X}{\Pr\left(T=c\right)}\\ &= \frac{1}{E\left[\frac{D}{P}\frac{Z_1-P}{1-P}\right]}\int E\left[Y^1-Y^0|X,T=c\right]E\left[\frac{D}{P}\frac{Z_1-P}{1-P}|X\right]dF_X\end{aligned}$$

,where we made use of (5) and a similar result for the fraction of compliers conditional on X. Now combining all results we obtain

$$E\left[Y^1-Y^0|T=c\right]$$
$$=\frac{1}{E\left[\frac{D}{P}\frac{Z_1-P}{1-P}\right]}\int\left(\int_0^1\frac{1}{\eta}\left(\ \Xi_1(X,\eta)-\Xi_0(X,\eta)\ \right)\frac{w(\eta,X)}{\int w(\eta,X)d\eta}\right)E\left[\frac{D}{P}\frac{Z_1-P}{1-P}|X\right]d\eta dF_X.$$

A.4 Influence function

Using the approach of Newey (2004), we obtain the influence function

$$
\begin{aligned}
& IF = YR \cdot \frac{Z_1 - \pi(X)}{\pi(X)\left(1-\pi(X)\right)} \cdot \left\{ D \frac{w(\Psi_1, X)}{\Psi_1 \cdot f\left(\Psi_1|X\right)} + (1-D) \frac{w(\Psi_0, X)}{\Psi_0 \cdot f\left(\Psi_0|X\right)} \right\} - \tau \\
+ \quad & E\left[YR \left\{ D \frac{w(\Psi_1, X)}{\Psi_1 \cdot f\left(\Psi_1|X\right)} + (1-D) \frac{w(\Psi_0, X)}{\Psi_0 \cdot f\left(\Psi_0|X\right)} \right\} \cdot (-1) \left\{ \frac{Z_1 - \pi(X)}{\pi(X)\left(1-\pi(X)\right)} \right\}^2 |X \right] \cdot \left(Z_1 - \pi(X)\right) \\
+ \quad & E\left[YR \frac{Z_1 - \pi(X)}{\pi(X)\left(1-\pi(X)\right)} D \frac{w(\Psi_1, X)}{\Psi_1 \cdot f\left(\Psi_1|X\right)} |X \right] - E\left[YR \frac{Z_1 - \pi(X)}{\pi(X)\left(1-\pi(X)\right)} D \frac{w(\Psi_1, X)}{\Psi_1 \cdot f\left(\Psi_1|X\right)} |\Psi_1, X \right] \\
+ \quad & E\left[YR \frac{Z_1 - \pi(X)}{\pi(X)\left(1-\pi(X)\right)} (1-D) \frac{w(\Psi_0, X)}{\Psi_0 \cdot f\left(\Psi_0|X\right)} |X \right] - E\left[YR \frac{Z_1 - \pi(X)}{\pi(X)\left(1-\pi(X)\right)} (1-D) \frac{w(\Psi_0, X)}{\Psi_0 \cdot f\left(\Psi_0|X\right)} |\Psi_0, X \right],
\end{aligned}
$$

where the second line represents the correction term for the nonparametric estimation of $\pi(X) = \Pr(Z_1 = 1|X)$, the third line that for the estimation of $f\left(\Psi_1|X\right)$, and the fourth line that for the estimation of $f\left(\Psi_0|X\right)$.

Calculating the expected square of this influence function led to a very lengthy expression containing many terms that include the weighting function w in non-linear ways. Minimizing this variance with respect to the choice of w in the class of non-zero functions integrating to one is non-trivial. Yet, even if one had obtained the optimal weighting function that minimizes the efficiency bound, it would contain many unknown conditional expectations involving Y, R, D, Z_1 and Ψ_d. Although all these conditional expectations can be estimated consistently nonparametrically, such estimates would be noisy in small samples and thus could lead to a noisy estimate of the weighting function, which could imply that some weights become arbitrarily large. Hence, the analytically optimal weighting function might behave poorly in finite samples and to guard against such poor behavior we would have to introduce a further trimming function on the *estimated* weighting function. We therefore seek a simpler (but yet intuitive) rule-of-thumb, which we develop in the following by only examining the first term of the influence function. In addition, the second rule-of-thumb we develop has the advantage that it does not involve the outcome data. This is a valuable property as it implies that the true (and unknown) treatment effects do not enter the calculation of the weighting function.[17]

Calculations with first term only

The first term of the influence function is

$$
IF^* = YR \cdot \frac{Z_1 - \pi(X)}{\pi(X)\left(1-\pi(X)\right)} \cdot \left\{ D \frac{w(\Psi_1, X)}{\Psi_1 \cdot f\left(\Psi_1|X\right)} + (1-D) \frac{w(\Psi_0, X)}{\Psi_0 \cdot f\left(\Psi_0|X\right)} \right\} - \tau.
$$

The semiparametric efficiency bound is given by the expected square of the influence function. Hence, pretending π and $f\left(\Psi_d|X\right)$ were not estimated, we obtain

$$
E\left[\left(IF^*\right)^2 \right] = E\left[\left(YR \cdot \frac{Z_1 - \pi(X)}{\pi(X)\left(1-\pi(X)\right)} \cdot \left\{ D \frac{w(\Psi_1, X)}{\Psi_1 \cdot f\left(\Psi_1|X\right)} + (1-D) \frac{w(\Psi_0, X)}{\Psi_0 \cdot f\left(\Psi_0|X\right)} \right\} - \tau \right)^2 \right],
$$

[17] This is somewhat akin to the approach in propensity score matching, where one can re-specify the propensity score for better balance, while being ensured that this specification process of the propensity score is not driven by the true treatment effects, thereby avoiding any (un)conscious data mining with respect to the outcome variable.

which we can re-write as

$$
\begin{aligned}
E\left[(IF^*)^2\right]-\tau^2 &= E\left[\left(YR\cdot\frac{Z_1-\pi(X)}{\pi(X)\left(1-\pi(X)\right)}\cdot\left\{D\frac{w(\Psi_1,X)}{\Psi_1\cdot f\left(\Psi_1|X\right)}+(1-D)\frac{w(\Psi_0,X)}{\Psi_0\cdot f\left(\Psi_0|X\right)}\right\}\right)^2\right] \\
&= E\left[Y^2R^2\cdot\left(\frac{Z_1-\pi(X)}{\pi(X)\left(1-\pi(X)\right)}\right)^2\cdot\left\{D\frac{w(\Psi_1,X)}{\Psi_1\cdot f\left(\Psi_1|X\right)}+(1-D)\frac{w(\Psi_0,X)}{\Psi_0\cdot f\left(\Psi_0|X\right)}\right\}^2\right] \\
&= E\left[Y^2R\cdot\left(\frac{Z_1}{\pi(X)^2}+\frac{1-Z_1}{\left(1-\pi(X)\right)^2}\right)\cdot\left\{D\frac{w^2(\Psi_1,X)}{\Psi_1^2\cdot f^2\left(\Psi_1|X\right)}+(1-D)\frac{w^2(\Psi_0,X)}{\Psi_0^2\cdot f^2\left(\Psi_0|X\right)}\right\}\right] \\
&= E\left[Y^2R\cdot\left(\frac{Z_1}{\pi(X)^2}+\frac{1-Z_1}{\left(1-\pi(X)\right)^2}\right)\cdot D\frac{w^2(\Psi_1,X)}{\Psi_1^2\cdot f^2\left(\Psi_1|X\right)}\right] \\
&\quad +E\left[Y^2R\cdot\left(\frac{Z_1}{\pi(X)^2}+\frac{1-Z_1}{\left(1-\pi(X)\right)^2}\right)\cdot(1-D)\frac{w^2(\Psi_0,X)}{\Psi_0^2\cdot f^2\left(\Psi_0|X\right)}\right] \\
&= \int E\left[Y^2RD\left(\frac{Z_1}{\pi(X)^2}+\frac{1-Z_1}{\left(1-\pi(X)\right)^2}\right)\frac{w^2(\Psi_1,X)}{\Psi_1^2\cdot f^2\left(\Psi_1|X\right)}|\Psi_1,X\right]f_{\Psi_1|X}\cdot d\Psi_1\cdot dF_X \\
&\quad +\int E\left[Y^2R(1-D)\left(\frac{Z_1}{\pi(X)^2}+\frac{1-Z_1}{\left(1-\pi(X)\right)^2}\right)\frac{w^2(\Psi_0,X)}{\Psi_0^2\cdot f^2\left(\Psi_0|X\right)}|\Psi_0,X\right]f_{\Psi_0|X}\cdot d\Psi_0\cdot dF_X \\
&= \int E\left[Y^2RD\left(\frac{Z_1}{\pi(X)^2}+\frac{1-Z_1}{\left(1-\pi(X)\right)^2}\right)|\Psi_1,X\right]\frac{w^2(\Psi_1,X)}{\Psi_1^2\cdot f\left(\Psi_1|X\right)}\cdot d\Psi_1\cdot dF_X \\
&\quad +\int E\left[Y^2R(1-D)\left(\frac{Z_1}{\pi(X)^2}+\frac{1-Z_1}{\left(1-\pi(X)\right)^2}\right)|\Psi_0,X\right]\frac{w^2(\Psi_0,X)}{\Psi_0^2\cdot f\left(\Psi_0|X\right)}\cdot d\Psi_0\cdot dF_X \\
&= \int\frac{w^2(\eta,X)}{\eta^2}\left\{\frac{\lambda_1(\eta,X)}{f_{\Psi_1|X}\left(\eta|X\right)}+\frac{\lambda_0(\eta,X)}{f_{\Psi_0|X}\left(\eta|X\right)}\right\}\cdot d\eta\cdot dF_X,
\end{aligned}
$$

where $\lambda_1(\eta,X)=E[Y^2RD\left(\frac{Z_1}{\pi(X)^2}+\frac{1-Z_1}{(1-\pi(X))^2}\right)|\Psi_1=\eta,X]$ and $\lambda_0(\eta,X)=E[Y^2R(1-D)\left(\frac{Z_1}{\pi(X)^2}+\frac{1-Z_1}{(1-\pi(X))^2}\right)|\Psi_0=\eta,X]$.