Received: 17 May 2019 | Revised: 11 March 2020

DOI: 10.1002/jae.2764

RESEARCH ARTICLE

WILEY Journal of APPLIED ECONOMETRICS

Endogeneity and non-response bias in treatment evaluation – nonparametric identification of causal effects by instruments

Hans Fricke[1] | **Markus Frölich**[2] | **Martin Huber**[3] | **Michael Lechner**[4]

[1] Amazon, San Francisco

[2] Department of Economics, University of Mannheim, Germany

[3] Department of Economics, University of Fribourg, Switzerland

[4] Department of Economics, University of St. Gallen, Switzerland

Correspondence
Martin Huber, University of Fribourg,
Bd. de Pérolles 90, CH-1700 Fribourg.
Email: martin.huber@unifr.ch

Summary

This paper proposes a nonparametric method for evaluating treatment effects in the presence of both treatment endogeneity and attrition/non-response bias, based on two instrumental variables. Using a discrete instrument for the treatment and an instrument with rich (in general continuous) support for non-response/attrition, we identify the average treatment effect on compliers as well as the total population under the assumption of additive separability of observed and unobserved variables affecting the outcome. We suggest non- and semiparametric estimators and apply the latter to assess the treatment effect of gym training, which is instrumented by a randomized cash incentive paid out conditional on visiting the gym, on self-assessed health among students at a Swiss university. The measurement of health is prone to non-response, which is instrumented by a cash lottery for participating in the follow-up survey.

1 | INTRODUCTION

The evaluation of the causal effect of a 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 different instruments in order to identify average treatment effects on treatment compliers (i.e. whose treatment complies with its instrument) as well as the total population. A crucial condition is additive separability of the observed and unobserved terms in the outcome equation, which implies constant treatment effects conditional on observed covariates across populations with different unobserved characteristics. This permits identifying treatment parameters in all of the population rather than e.g. exclusively among those always responding under either treatment, a group considered for instance in Imai (2008) and Lee (2009), or always responding compliers, see Chen and Flores (2015).[1]

[1] Discussing nonparametric point identification based on distinct instruments for the treatment assignment and attrition appears related to Lee and Salanié (2018), who show nonparametric identification under multiple or dynamic endogenous treatments based on multiple instruments. However, the identifying assumptions differ when applying their framework to the case of outcome attrition and treatment endogeneity, which would not allow for an effect of the treatment on attrition. Furthermore, Lee and Salanié (2018) assume all instruments to be continuous, but do not require additive separability in the outcome equation. In contrast, in this paper the treatment may affect attrition and the treatment's instrument may be binary, but additive separability needs to hold. We are grateful to an anonymous referee for pointing this out.

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 could be generalized to multi-valued instruments. Concerning endogenous outcome non-response, we assume an instrument with rich enough support: Conditional on observed covariates, the instrument must affect response in a way that there exist treated and non-treated compliers with the same, non-zero response probability. In general, this requires the instrument to be continuous. The method appears for instance useful for tackling attrition in follow-up surveys (be it online, mail, or phone surveys or face-to-face interviews) when using financial incentives such as vouchers, cash payments, or cash lotteries as instruments, see e.g. Castiglioni, Pforr, and Krieger (2008), Pforr et al. (2015), and Hsu, Schmeiser, Haggerty, and Nelson (2017). Discounts on products or services as well as charitable donations are further incentives that might possibly be exploited as continuous instruments for non-response. To increase statistical power, they could be combined with non-continuous instruments like the number of reminders to complete a follow-up survey (e.g. e-mail/mail reminders or phone calls). 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 non- and semiparametric estimators based on regression and weighting. We also provide a simulation study that suggests that the regression-based estimators perform well in samples with a few 1000 observations.

We apply our method to evaluate the effect of students' physical (gym) training on self-assessed health at the University of St. Gallen in Switzerland. The application is rather unique in the sense that it contains two separately randomized 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. These studies do not consider the problem of outcome non-response or attrition. In the presence of the latter, treatment effects are generally not point identified. See Chen and Flores (2015), Blanco, Chen, Flores, and Flores-Lagunes (2016), and Dong (2019), who discuss set identification of treatment effects (i.e. derive bounds on the latter) under treatment endogeneity and attrition when only a single instrument for the treatment is available.

Void of a second instrument for attrition, point identification requires the imposition of stringent structural assumptions on the attrition process. One popular assumption is the so-called missing at random (MAR) restriction, which imposes conditional exogeneity of attrition (with respect to potential outcomes) 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. An alternative to MAR which is applicable to the LATE framework is the latent ignorability (LI) assumption, see Frangakis and Rubin (1999). 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. In such parametric models, instruments for non-ignorable non-response may nevertheless be useful for preventing multicollinearity problems, while they are a precondition for identification in nonparametric models. See for instance Huber (2012,2014), who considers a continuous instrument for attrition when evaluating a treatment that is random or conditionally exogenous given observables. This permits identifying treatment effects in the total population (rather than among compliers alone as in the endogenous treatment context of this paper) under additive separability of observed and unobserved terms in the outcome equation, or in the subpopulation of respondents if additivity does not hold.

Also Behaghel, Crépon, Gurgand, and Le Barbanchon (2015) consider a nonadditive outcome model with a randomized treatment, but predominantly focus on a discrete (rather than continuous) instrument for response. They provide conditions for point identifying treatment effects among (a subset of) respondents, as well as for deriving informative bounds (that are tighter than e.g. those in Lee, 2009) and to this end assume monotonicity of response in its instrument. 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.[2]

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 et al. (2015) use phone calls as (quasi-)instrument for non-response to bound 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 et al., 2006), while more flexible (nonparametric) modelling only allows for a partial identification of the treatment effect (see Behaghel et al., 2015). 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. The proposed methods permit designing 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 nonparametric identification approaches. Section 4 discusses estimation. Section 5 presents simulation evidence on the finite sample properties of the estimation approach. Section 6 discusses the application to an 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 is 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:

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

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

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

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

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

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

Frölich (2007). Z_2 is the instrument for response R, henceforth referred to as second instrument, which is assumed to be of rich, in general of *continuous* support. In our model, we permit the two instruments to be correlated and the compliance type to be correlated with Z_2. However, we require Z_1 to be excluded from the response Equation (2), such that it must not influence R other than via its effect on the treatment. In the context of our application of Section 6, this rules out that the mere offer of a cash incentive to exercise (Z_1) directly affects response behavior, i.e. independent of actually visiting the gym (D) upon which the incentive's payout was conditional. This would e.g. be violated if offering the incentive induced a feeling of gratitude that directly increased the response rate or if not offering it created disappointment reducing the inclination to respond.[3]

Our model imposes additive separability in the outcome equation (1). This structure implies a conditionally constant treatment effect, i.e. that the treatment effect is identical for all individuals with the same values in X. While this permits arbitrary effect heterogeneity across X because function ϕ is completely unrestricted, it rules out heterogeneity in unobservables given X. Whenever treatment effects in the total rather than the responding population are of interest, additive separability allows weakening the support requirements on the instrument Z_2. One 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.[4] However, in many applications including ours, the instrument Z_2 is not that strong and therefore, non-response does not fully vanish under any value z_2 in the support of Z_2. Additive separability permits extrapolating treatment effects to such groups never responding due to 'too extreme' values in V. Analogously, this assumption also enables the extrapolation of the LATE among treatment compliers to other groups like the total population, see also Angrist and Fernández-Val (2010) and Aronow and Carnegie (2013). Without additive separability, identification would in general only be feasible in the subpopulation of always responding treatment compliers e.g. considered in Chen and Flores (2015) for bounding treatment effects in the absence of an instrument for response. The reason is that only for this subpopulation, potential outcomes are observed under either treatment state and can thus serve as base for any effect identification and extrapolation methods as proposed in this paper.

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:

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

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 et al. (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. $D_i(0) = 1, D_i(1) = 0$.

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. $E[Y^1 - Y^0 | T = c]$ 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] If Z_1 directly affects R such that the exclusion restriction is violated, point identification of treatment effects is generally lost because the distribution of V given X among compliers cannot be recovered, as required for our identification results presented in Theorem 1 further below.
[4] This has been referred to as 'identification at infinity' in the parametric literature on attrition, non-response and selection models.

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 associated with 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 condition (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 or experiments where randomization is within strata defined on X, on the other hand, Assumption 1 is often only plausible after conditioning on covariates X.

In addition to the independence assumptions, identification relies on the weak monotonicity of the treatment in its instrument. For this reason, Assumption 2 rules out the existence of defiers and invokes the existence of compliers for any value of X and Z_2 in their support.[5] If compliers existed only for a subset of values in the support, effects on the total population would not be identified, i.e. the LATE could not be extrapolated to the ATE. If Z_2 is randomly assigned and thus not associated with treatment compliance conditional on X as in condition (4), then the existence of compliers given X implies their existence conditional on both X and Z_2. Formally, $\Pr(T = c|\, X) > 0$ entails $\Pr(T = c|\, Z_2, X) > 0$ if $Z_2 \perp\!\!\!\perp T|X$. Concerning the number of elements in X, there exists a potential trade-off between the satisfaction of additive separability in model (1) and Assumption 2. While more covariates may make additive separabiliy (and thus effect heterogeneity in observables only) more plausible, the existence of compliers for any possible combination of covariates could become less likely. As $\Pr(T = c|Z_2, X)$ is identified further below, the existence of compliers given X and Z_2 can be tested in the data.

[5] 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 the first part of Assumption 2, i.e. $\Pr(T = c|\, Z_2, X) > 0$ is directly testable if the second part holds, i.e. $\Pr(T = d|\, Z_2, X) = 0$. The latter is, however, not directly testable, although see Huber and Mellace (2015), Kitagawa (2015), Mourifié and Wan (2017), and Machado, Shaikh, and Vytlacil (2019) for methods jointly testing monotonicity *and* IV independence.

Assumption 2. Weak monotonicity of treatment choice

$$\Pr(T = c|\, Z_2, X) > 0 \text{ and } \Pr(T = d|\, Z_2, X) = 0.$$

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

$$0 < \Pi < 1 \text{ and } 0 < P < 1 \text{ almost surely.}$$

It follows from Assumptions 1 to 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

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

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:***

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

Since the left hand side is identified, see definition (6), the conditional 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 plausible in many 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 $\zeta(1, z_2', x) = \zeta(0, z_2'', x)$. This result is crucial for identification, which is based on the following intuition. The sample selection/non-response problem occurs because we observe outcomes only for observations with $V \le \zeta(D, Z_2, X)$. 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 $\zeta(1, z_2', x) = \zeta(0, z_2'', x) = v$, then the set of observations with outcome data is given by $V \le v$ 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

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

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

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

Appendix A.3 demonstrates that the previous expressions may be reformulated as follows:

$$\frac{E\left[YR(Z_1 - p(z_2', x))|D = 1, X = x, Z_2 = z_2'\right]}{E\left[Z_1 - p(z_2', x)|D = 1, X = x, Z_2 = z_2'\right]} - \frac{E\left[YR(Z_1 - p(z_2'', x))|D = 0, X = x, Z_2 = z_2''\right]}{E\left[Z_1 - p(z_2'', x)|D = 0, X = x, Z_2 = z_2''\right]} \tag{8}$$

$$= E\left[\{\phi(1,x) + U\} \cdot 1\left\{\zeta(1, z_2', x) \geq V\right\} - \{\phi(0,x) + U\} \cdot 1\left\{\zeta(0, z_2'', x) \geq V\right\} |X = x, T = c\right]$$

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, the conditional treatment effect among compliers given X times the rank $\psi_1(z_2', x)$ is given by

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

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

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

For obtaining the LATE or ATE, we need to identify $E[Y^1 - Y^0|X, T = c]$ at almost every x in the complier or total population, respectively. This requires that for every x some values z_2' and z_2'' exist that satisfy $\psi_1(z_2', x) = \psi_0(z_2'', x)$. Formally, let $Supp\,(\Psi_1|X = x)$ 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

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

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 the identification of the conditional treatment effect almost everywhere. Common support in general requires Z_2 to be continuous, even though it can hold in specific cases even under a discrete Z_2. As one (admittedly very particular) example, assume a binary Z_2 and that D and Z_2 have exactly offsetting effects on Ψ_d given any value of X.

For every x, it is under additive separability in model (1) sufficient for identification to just choose one value in $\mathcal{X}_x$ and apply (9).[6]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

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

Since the previous result holds for any η, we may 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 an arbitrary weighting function $w(\eta, x)$ as a function of η and possibly also of x. The conditional treatment effect is given by

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

[6] This suggests a specification test based on testing whether conditional treatment effects differ across values in $\mathcal{X}_x$.

provided that the weighting function $w(\eta, x)$ does not integrate to zero. Therefore, integration over X among compliers gives the LATE, denoted by $E\left[Y^1 - Y^0 | T = c\right]$. 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), the ATE, denoted by $E\left[Y^1 - Y^0\right]$, is analogously obtained by integration over X in the total population. More generally and related to the weighted treatment effect in Hirano, Imbens, and Ridder (2003), the ATE on any target population defined in terms of the covariate distribution is identified by appropriately weighting the conditional effect (12) with respect to covariates X, see our main identification result in Theorem 1. This includes the average treatment effect on the treated (ATET), given by $E\left[Y^1 - Y^0 | D = 1\right]$. To ease notation, denote the integral of the weighting function for some value x as

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

and suppose that $c(x)$ is non-zero for almost every x. Furthermore, assume that $g(X)$ is a well behaved function satisfying that $|g(X)|$ is bounded from above and $E[g(X)] > 0$.

Theorem 1. *Under Assumptions 1 to 5,*

$$\begin{aligned}
E_{g(X)}\left[Y^1 - Y^0\right] &= \frac{1}{E[g(X)]} \int \int_0^1 \left(\Xi_1(X, \eta) - \Xi_0(X, \eta)\right) \frac{w(\eta, X)}{\eta \cdot c(X)} g(X) d\eta dF_X, \textit{with} \\
E_{g(X)}\left[Y^1 - Y^0\right] &= E\left[Y^1 - Y^0\right] \textit{ if } g(X) = E[g(X)] = 1, \\
E_{g(X)}\left[Y^1 - Y^0\right] &= E\left[Y^1 - Y^0 \middle| T = c\right] \textit{ if } g(X) = E\left[\frac{D}{P} \frac{Z_1 - P}{1 - P} \middle| X\right], E[g(X)] = E\left[\frac{D}{P} \frac{Z_1 - P}{1 - P}\right], \\
E_{g(X)}\left[Y^1 - Y^0\right] &= E\left[Y^1 - Y^0 | D = 1\right] \textit{ if } g(X) = \Pr(D = 1 | X), E[g(X)] = \Pr(D = 1).
\end{aligned}$$

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 $\Pr(Z_1 = 1 | Z_2, X, \Psi_1) = \Pr(Z_1 = 1 | Z_2, X) = \Pr(Z_1 = 1 | X)$ and therefore, $P = \Pi$. This also implies that D is independent of Ψ_d given X.[7] The control function thus simplifies to

$$\psi_d(z_2, x) \equiv \frac{E\left[R\left(Z_1 - \pi(x)\right) | D = d, X = x, Z_2 = z_2\right]}{E\left[Z_1 - \pi(x) | D = d, X = x, Z_2 = z_2\right]}. \tag{14}$$

We also note that $E\left[\frac{D}{\Pi} \frac{Z_1 - \Pi}{1 - \Pi} | X = x, \Psi_d = \eta\right] = E\left[\frac{D}{\Pi} \frac{Z_1 - \Pi}{1 - \Pi} | X = x\right] = \Pr(T = c | X = x)$ and that the expression of Theorem 1 simplifies considerably. See the identification result Lemma 2, where it is also shown that treatment effects are identified by a weighting expression in the spirit of inverse probability weighting (IPW, see the seminal work of Horvitz and Thompson (1952)), in which $f(\Psi_1 | X)$ and $f(\Psi_0 | X)$ denote the conditional densities of Ψ_1 and Ψ_0 given X.

Lemma 2. *Under Assumptions 1 to 5 and $Z_2 \perp\!\!\!\perp (Z_1, T) | X$,*

$$\begin{aligned}
E_{g(X)}\left[Y^1 - Y^0\right] &= \frac{1}{E[g(X)]} \int_0^1 \left(\frac{E\left[\frac{YRD}{\Pi} \frac{Z_1 - \Pi}{1 - \Pi} | X, \Psi_1 = \eta\right] + E\left[\frac{YR(1 - D)}{\Pi} \frac{Z_1 - \Pi}{1 - \Pi} | X, \Psi_0 = \eta\right]}{E\left[\frac{D}{\Pi} \frac{Z_1 - \Pi}{1 - \Pi} | X = x\right]} \right) \frac{w(\eta, X)}{\eta \cdot c(X)} g(X) d\eta dF_X \\
&= \frac{1}{E[g(X)]} E\left[\frac{YR}{E\left[\frac{D}{\Pi} \frac{Z_1 - \Pi}{1 - \Pi} | X = x\right]} \frac{Z_1 - \Pi}{\Pi(1 - \Pi) \cdot c(X)} \cdot \left\{ D \frac{w(\Psi_1, X)}{\Psi_1 \cdot f(\Psi_1 | X)} + (1 - D) \frac{w(\Psi_0, X)}{\Psi_0 \cdot f(\Psi_0 | X)} \right\} g(X) \right],
\end{aligned} \tag{15}$$

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

with

$$E_{g(X)}\left[Y^1 - Y^0\right] = E\left[Y^1 - Y^0\right] \textit{ if } g(X) = E[g(X)] = 1,$$

$$E_{g(X)}\left[Y^1 - Y^0\right] = E\left[Y^1 - Y^0 \middle| T = c\right] \textit{if } g(X) = E\left[\frac{D}{\Pi}\frac{Z_1 - \Pi}{1 - \Pi}|X = x\right], E[g(X)] = E\left[\frac{D}{\Pi}\frac{Z_1 - \Pi}{1 - \Pi}\right],$$

$$E_{g(X)}\left[Y^1 - Y^0\right] = E\left[Y^1 - Y^0|D = 1\right] \textit{ if } g(X) = \Pr(D = 1|X), E[g(X)] = \Pr(D = 1).$$

Natural estimators follow from replacing unconditional expectations in the latter equations by sample means and plugging in nonparametric estimators of the other components, see Eqs. (20), (21), (22), and (23) 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 (15) with $g(X) = E\left[\frac{D}{\Pi}\frac{Z_1-\Pi}{1-\Pi}|X = x\right]$ for LATE estimation under a given weighting function $w(\cdot)$. The results analogously apply to 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 weighting function. For this reason, we derive the semiparametric efficiency bound of

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

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 Appendix A.3, we derive the 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 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

$$\lambda_1(\eta, X) = E[Y^2 RD\left(\frac{Z_1}{\pi(X)^2} + \frac{1 - Z_1}{\left(1 - \pi(X)\right)^2}\right)|\Psi_1 = \eta, X], \tag{16}$$

$$\lambda_0(\eta,X)=E[Y^2R(1-D)\left(\frac{Z_1}{\pi(X)^2}+\frac{1-Z_1}{(1-\pi(X))^2}\right)|\Psi_0=\eta,X]. \tag{17}$$

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

Weighting function 1:

$$w(\eta,x)\propto\frac{\eta}{\sqrt{\frac{\lambda_1(\eta,x)}{f_{\Psi_1|X=x}(\eta|x)}+\frac{\lambda_0(\eta,x)}{f_{\Psi_0|X=x}(\eta|x)}}}. \tag{18}$$

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:

$$w(\eta,x)\propto\frac{\eta}{\sqrt{\frac{1}{f_{\Psi_1|X=x}(\eta|x)}+\frac{1}{f_{\Psi_0|X=x}(\eta|x)}}}. \tag{19}$$

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 with $g(X)=E\left[\frac{D}{P}\frac{Z_1-P}{1-P}|X\right]=\Pr(T=c|X)$ 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.

$$\begin{aligned}\widehat{LATE}=&\frac{1}{\hat{\Pr}(T=c)}\frac{1}{n}\sum_{i=1}^{n}\sum_{\eta\in\mathcal{S}}\frac{\hat{w}(\eta,X_i)}{\sum_{\eta\in\mathcal{S}}\hat{w}(\eta,X_i)}\frac{1}{\eta}\\&\times\left(\frac{\hat{E}[YRD(Z_1-\hat{\pi}(X))|\hat{\Psi}_1=\eta,X_i]+\hat{E}[YR(1-D)(Z_1-\hat{\pi}(X))|\hat{\Psi}_0=\eta,X_i]}{\hat{\pi}(X_i)(1-\hat{\pi}(X_i))}\right),\end{aligned} \tag{20}$$

where

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

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 (14), 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. We compute $\hat{E}[Y_iR_iD_i(Z_{1i}-\hat{\pi}(X_i))|\eta,X_i]$ and $\hat{E}[Y_iR_i(1-D_i)(Z_{1i}-\hat{\pi}(X_i))|\eta,X_i]$ in (20) by a local linear regression of $Y_iR_iD_i(Z_{1i}-\hat{\pi}(X_i))$ and $Y_iR_i(1-D_i)(Z_{1i}-\hat{\pi}(X_i))$ 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 (18) or (19). For the first weighting approach, see (18), we also require estimates of λ_1 and λ_0, which we obtain by local linear regression of $Y_i^2R_iD_i[Z_{1i}/\hat{\pi}(X_i)^2+(1-Z_{1i})/(1-\hat{\pi}(X_i))^2]$ on $(1,\hat{\Psi}_{i1},X_i)$ and $Y_i^2R_i(1-D_i)[Z_{1i}/\hat{\pi}(X_i)^2+(1-Z_{1i})/(1-\hat{\pi}(X_i))^2]$ on $(1,\hat{\Psi}_{i0},X_i)$ to estimate equations (16) and (17). Finally, $\mathcal{S}$ denotes the set of support points of η considered in our LATE estimator (20) which approximates the integral over η in Lemma 2. In our simulations and application, it consists of an equidistant 100-points grid of values with the smallest value being the maximum out of the minimum of $\hat{\Psi}_{i1}$, the minimum of $\hat{\Psi}_{i0}$, and 0.01. The last (and largest) value in the grid is the minimum out of the maximum of $\hat{\Psi}_{i1}$, the maximum of $\hat{\Psi}_{i0}$, and 1.[8]

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

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).[9]

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.

The ATE can be estimated in analogy to the LATE, when using Lemma 2 with $g(X) = 1$ as weighting function of the covariates:

$$\begin{aligned} A\hat{T}E = & \frac{1}{n}\sum_{i=1}^{n}\sum_{\eta\in S}\frac{\hat{w}(\eta, X_i)}{\sum_{\eta\in S}\hat{w}(\eta, X_i)}\frac{1}{\eta} \\ & \times\left(\frac{\hat{\mathrm{E}}[YRD(Z_1-\hat{\pi}(X))|\hat{\Psi}_1=\eta, X_i]+\hat{\mathrm{E}}[YR(1-D)(Z_1-\hat{\pi}(X))|\hat{\Psi}_0=\eta, X_i]}{\hat{\mathrm{E}}\,[DZ_1|X_i]-\hat{\mathrm{E}}\,[D|X_i]\,\hat{\pi}(X_i)}\right), \end{aligned} \tag{21}$$

with $\hat{\mathrm{E}}\,[DZ_1|X_i]$ and $\hat{\mathrm{E}}\,[D|X_i]$ denoting estimates of $E\,[DZ_1|X_i]$ and $E\,[D|X_i]$. In addition to the regression-based estimator in (20), Lemma 2 suggests the following IPW estimator of the LATE:

$$\begin{aligned} LA\hat{T}E = & \frac{1}{\hat{\Pr}(T=c)}\frac{1}{n}\sum_{n}^{i=1}Y_iR_i\frac{Z_{1i}-\hat{\pi}(X_i)}{\hat{\pi}(X_i)\,(1-\hat{\pi}(X_i))\cdot c(X_i)} \\ & \times\left\{D_i\frac{w(\hat{\Psi}_{1i}, X_i)}{\hat{\Psi}_{1i}\cdot\hat{f}\left(\hat{\Psi}_{1i}|X_i\right)}+(1-D_i)\frac{w(\hat{\Psi}_{0i}, X_i)}{\hat{\Psi}_{0i}\cdot\hat{f}\left(\hat{\Psi}_{0i}|X_i\right)}\right\}, \end{aligned} \tag{22}$$

where $c(X_i)$ captures the scaling of the weighting function given by (13). Analogously, an IPW estimator for the ATE is obtained by

$$\begin{aligned} A\hat{T}E = & \frac{1}{n}\sum_{n}^{i=1}Y_iR_i\frac{Z_{1i}-\hat{\pi}(X_i)}{\hat{\mathrm{E}}\,[DZ_1|X_i]-\hat{\mathrm{E}}\,[D|X_i]\,\hat{\pi}(X_i)}\frac{1}{c(X_i)} \\ & \times\left\{D_i\frac{w(\hat{\Psi}_{1i}, X_i)}{\hat{\Psi}_{1i}\cdot\hat{f}\left(\hat{\Psi}_{1i}|X_i\right)}+(1-D_i)\frac{w(\hat{\Psi}_{0i}, X_i)}{\hat{\Psi}_{0i}\cdot\hat{f}\left(\hat{\Psi}_{0i}|X_i\right)}\right\}, \end{aligned} \tag{23}$$

with $\hat{\mathrm{E}}\,[DZ_1|X_i]$ and $\hat{\mathrm{E}}\,[D|X_i]$ denoting estimates of $E\,[DZ_1|X_i]$ and $E\,[D|X_i]$. 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 (20), we conduct a simulation study based on the following data generating process (DGP):

$$Y_i = D_i - 0.5X_i + U_i, \tag{24}$$

$$\begin{aligned} & Y_i \text{ is observed if } R_i = 1, \\ & R_i = I\{D_i - 0.5X_i + Z_{2i} + V_i > 0\}, \end{aligned} \tag{25}$$

$$D_i = I\{Z_{1i} - 0.5X_i + W_i > 0\}, \tag{26}$$

$$Z_{1i} = I\{-0.5X_i + P_i > 0\}, \tag{27}$$

$$Z_{2i} = -0.5X_i + Q_i, \tag{28}$$

[9] Using cross-validated bandwidths did not importantly affect the results of the application.

$$P_i, Q_i \sim \mathcal{N}(0,1), \text{independently of each other and of}(X_i, U_i, V_i, W_i),$$
$$\begin{pmatrix} U_i \\ V_i \\ W_i \end{pmatrix} \sim \mathcal{N}(\mu, \Sigma), \text{where } \mu = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \text{and } \Sigma = \begin{pmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{pmatrix} \tag{29}$$

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 (18) and (19) are considered as weighting functions in estimation based on (20). 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 nonpar IV' and 'LATE semipar IV', respectively. As a comparison, we also include a LATE estimator that is consistent under MAR, i.e. when response is selective with respect to observables Z_1, X alone. The method is related to the IPW approach in Frölich (2007) based on weighting by the inverse of the instrument propensity $\Pr(Z_1 = 1|X)$. As an important difference, however, outcomes are additionally weighted by the inverse of the response propensity $\Pr(R = 1|Z_1, X)$ to adjust for non-response associated with observables as for instance also considered in Schochet, Burghardt, and McConnell (2000). More formally, we consider the sample analogue of the following IPW expression:

$$E\left[\frac{YR}{\Pr(R=1|Z_1,X)\Pi}\frac{Z_1-\Pi}{1-\Pi}\right] \Big/ E\left[\frac{D}{\Pi}\frac{Z_1-\Pi}{1-\Pi}\right], \tag{30}$$

which can be shown to correspond to $\frac{E\left[E[Y|R=1,Z_1=1,X]-E[Y|R=1,Z_1=0,X]\right]}{E\left[E[D|Z_1=1,X]-E[D|Z_1=0,X]\right]}$ by the law of iterated expectations and basic probability theory. If MAR is satisfied, $E[Y|R = 1, Z_1, X] = E[Y|Z_1, X]$ such that (30) then corresponds to the LATE if Z_1 is a valid instrument, see the discussion in Frölich (2007). We estimate $\Pr(R = 1|Z_1, X)$ and Π either nonparametrically ('LATE nonpar MAR') or by probit ('LATE semipar MAR'). We consider two sample sizes ($n = 1000, 4000$) and run 1000 simulations with various specifications.

Table 1 presents the bias, standard deviation ('st.dev.'), and root mean squared error ('RMSE') of the estimators when the observed confounder is binary, i.e. $X_i \sim$ binom(0.5). The complier and response rates are 36% and 51%, respectively, under this covariate distribution. Nonparametric estimation with instrument-based non-response correction ('LATE nonpar IV') 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. We generally find that the simpler weighting function (19), where fewer components need to be estimated, performs better. Precision and RMSE improve under the larger sample size and when using weighting function (19), the nonparametric estimator now outperforms the semi- and nonparametric LATE estimators assuming MAR, which is severely biased due to omitting non-response related to unobservables. However, semiparametric estimation with instrument-based non-response correction ('LATE semipar IV') performs considerably better with respect to precision and RMSE than the nonparametric method under

TABLE 1 Simulations with binary covariate

	***n*=1000**			***n*=4000**		
	bias	**st.dev.**	**RMSE**	**bias**	**st.dev.**	**RMSE**
LATE nonpar IV w1	−0.06	1.75	1.75	−0.01	0.52	0.52
LATE nonpar IV w2	0.01	0.58	0.58	0.00	0.21	0.21
LATE semipar IV w1	−0.01	0.31	0.31	−0.01	0.16	0.16
LATE semipar IV w2	0.00	0.32	0.32	−0.00	0.16	0.16
LATE nonpar MAR	−0.49	0.28	0.57	−0.49	0.14	0.51
LATE semipar MAR	−0.49	0.28	0.57	−0.49	0.14	0.51

Note: 'st.dev.' denotes the standard deviation, 'RMSE' the root mean squared error of the respective estimator. 'LATE nonpar IV', 'LATE semipar IV', 'LATE nonpar MAR' and 'LATE semipar MAR' refers to nonparametric estimation based on (20), semiparametric estimation based on (20) with parametric first step estimators, and non- and semiparametric LATE estimation under MAR based on (30), respectively. 'w1' and 'w2' stands for weighting based on (18) and (19), respectively.

TABLE 2 Simulations with continuous covariate

	n=1000			n=4000		
	bias	**st.dev.**	**RMSE**	**bias**	**st.dev.**	**RMSE**
LATE nonpar IV w1	−0.00	0.75	0.75	−0.03	0.23	0.23
LATE nonpar IV w2	−0.06	0.54	0.55	−0.04	0.22	0.22
LATE semipar IV w1	0.01	0.32	0.32	0.00	0.16	0.16
LATE semipar IV w2	0.02	0.32	0.32	0.01	0.17	0.17
LATE nonpar MAR	−0.50	0.24	0.55	−0.49	0.12	0.51
LATE semipar MAR	−0.50	0.24	0.56	−0.49	0.12	0.51

Note: 'st.dev.' denotes the standard deviation, 'RMSE' the root mean squared error of the respective estimator. 'LATE nonpar IV', 'LATE semipar IV', 'LATE nonpar MAR' and 'LATE semipar MAR' refers to nonparametric estimation based on (20), semiparametric estimation based on (20) with parametric first step estimators, and non- and semiparametric LATE estimation under MAR based on (30), respectively. 'w1' and 'w2' stands for weighting based on (18) and (19), respectively.

TABLE 3 Simulations with quadratic relation

	n=1000			n=4000		
	bias	**st.dev.**	**RMSE**			
LATE nonpar IV w1	0.05	0.42	0.42	0.01	0.19	0.19
LATE nonpar IV w2	0.02	0.43	0.43	−0.00	0.19	0.19
LATE semipar IV w1	−0.03	0.32	0.33	0.00	0.16	0.16
LATE semipar IV w2	−0.00	0.33	0.33	−0.00	0.17	0.17
LATE nonpar MAR	−0.25	0.23	0.35	−0.26	0.12	0.29
LATE semipar MAR	−0.25	0.23	0.35	−0.26	0.12	0.28

Note: 'st.dev.' denotes the standard deviation, 'RMSE' the root mean squared error of the respective estimator. 'LATE nonpar IV', 'LATE semipar IV', 'LATE nonpar MAR' and 'LATE semipar MAR' refers to nonparametric estimation based on (20), semiparametric estimation based on (20) with parametric first step estimators, and non- and semiparametric LATE estimation under MAR based on (30), respectively. 'w1' and 'w2' stands for weighting based on (18) and (19), respectively.

either sample size. It even dominates LATE estimation assuming MAR under $n = 1000$, depends little on the chosen weighting function, and appears to be the preferred choice in smaller samples.

Table 2 reports the results when the covariate follows a uniform distribution between -0.5 and 0.5, i.e. $X_i \sim \mathcal{U}[-0.5, 0.5]$. The complier and response rates are 34% and 66%, respectively. Again, fully nonparametric estimation with instrument-based non-response correction 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 with instrument-based non-response correction. The latter method entails RMSEs that are similar to those in Table 1 and is again stable across weighting schemes. LATE estimation assuming MAR 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.

The previous simulation scenarios arguably favored the semiparametric methods relying on parametric first step estimators, because the outcome was modeled linearly, while treatment and response corresponded to probit models with a linear index. In our third specification, we therefore introduce nonlinearities in the equations violating these specifications. To this end, we replace $-0.5X_i$ in equations (24) to (28) by X_i^2, change the distribution of the covariate to $X_i \sim \mathcal{U}[0, 0.5]$, and replace the 0 in equation (25) by 0.2. Furthermore, we set the effect of D_i on Y_i in (24) to X_i^2, implying treatment effect heterogeneity with respect to the observable. The complier and response rates are 33% and 67%, respectively. Table 3 reports the results. In spite of the misspecification of the parametric first step estimators, the semiparametric methods with instrument-based non-response correction remain competitive. Their bias is not larger than that of the nonparametric approaches, however, their variance is somewhat smaller. LATE estimation assuming MAR has the smallest variance, but is prone to a non-negligible bias and thus becomes relatively less attractive as the sample size increases.

Finally, Table 4 considers a simulation scenario that more closely matches our empirical application in Section 6. First, we use a sample size of $n = 500$ observations. Second, the effect of D_i in outcome equation (24) with the continuous covariate ($X_i \sim \mathcal{U}[-0.5, 0.5]$) is set to 0.6 (rather than 1) and the continuous outcome Y_i is transformed into a discrete one taking integer values between 1 and 5 using the 'draw_likert' command of the 'fabricatr' package for R. Third, the coefficient on Z_{1i} in treatment equation (26) is set to 0.3 (rather than 1). All remaining parameters are the same as in the

TABLE 4 Simulations with continuous covariate, n=500, and discrete outcome

	bias	st.dev.	RMSE
LATE nonpar IV w1	0.38	12.76	12.77
LATE nonpar IV w2	−0.54	10.88	10.90
LATE semipar IV w1	0.49	6.03	6.05
LATE semipar IV w2	0.63	6.08	6.11
LATE nonpar MAR	−0.45	2.69	2.72
LATE semipar MAR	−0.37	1.87	1.90

Note: 'st.dev.' denotes the standard deviation, 'RMSE' the root mean squared error of the respective estimator. 'LATE nonpar IV', 'LATE semipar IV', 'LATE nonpar MAR' and 'LATE semipar MAR' refers to nonparametric estimation based on (20), semiparametric estimation based on (20) with parametric first step estimators, and non- and semi-parametric LATE estimation under MAR based on (30), respectively. 'w1' and 'w2' stands for weighting based on (18) and (19), respectively.

first and second simulation design. These settings entail a true LATE of 0.59, a complier share of 12%, and a response rate 63%, parameters that are by and large comparable to the estimates of the empirical application. Table 4 reports the results. Due to the moderate sample size and small complier share, the performance of any estimator is considerably poorer than in the previous simulations. In terms of bias, the methods with instrument-based non-response correction are no longer preferable to LATE assuming MAR. Furthermore, the standard deviations of both our semi- and nonparametric versions explode when compared to the previous scenarios and are even considerably larger than the already substantial standard deviation of LATE assuming MAR. The latter thus dominates in terms of having the smallest RMSE. However, the performance of any method appears unsatisfactory. Therefore, the empirical application further below should be merely regarded as an illustration for the kind of designs the proposed estimators may be applied to, even though a considerably larger sample would be required for sufficiently precise inference.

6 | APPLICATION: THE EFFECTS OF SPORTS ON SELF-REPORTED HEALTH

6.1 | The experiment

The estimators of the LATE and ATE outlined in (20) and (21), respectively, are applied in a field experiment to analyze the short-term effect of recreational sport and exercise in university on self-assessed health.[10] 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. These amenities are costly and often compete with funds for other activities such as teaching or research. For the US, Jacob, McCall, and Stange (2018) document that non-profit 4-year colleges in the US spend on average 50 cents on recreational amenities for each dollar spent on academics. Therefore, it is important to know whether participation in these amenities improves health outcomes for students as intended by universities. While in general health benefits of sports and physical exercise are well established,[11] 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 characteristics from those students that do not.[12] To solve this endogeneity problem, Fricke, Lechner, and Steinmayr (2015) carried out an experiment at the University of St.Gallen,[13] 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 randomly with cash incentives to participate in campus sports and exercise.[14] Half of the students received a cash incentive of 100 CHF (approximately worth USD 110 at that time), while the other half did not. The exact implementation was as follows: Students were split into 13 blocks conditional on individual characteristics. In all blocks, approximately half of the students were assigned to the treatment

[10] We acknowledge that self-assessed health could be measured with error. The identification of treatment effects then requires the measurement error in the outcome to be unrelated to the treatment among compliers, conditional on covariates (X) and the conditional response probability (Ψ_d). Otherwise, point identification is generally lost.

[11] SeeTimmons et al. (2012) for small children, Janssen and Leblanc (2010) for adolescents, and Reiner, Niermann, Jekauc, and Woll (2013) for adults.

[12] See for exampleSchneider and Becker (2005), and Farrell and Shields (2002).

[13] 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.

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

group or control group. If students used the campus sports facilities twice per week over ten weeks, they received the entire amount. Each week the endowment was reduced by CHF 5 if they participated only once a week, or by CHF 10 if they did not participate at all.

Using this experiment, we randomly invited 80% of the students to join a lottery in which they could win a cash price with a chance of 25%, conditional on participation in a follow-up survey measuring self-reported health. The cash prices varied in steps of CHF 10 from 10 to CHF 200 to incentivize participation and again, randomization was within three blocks conditional on individual characteristics. 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 emails offering the cash lottery conditional on survey participation after the first survey email and after the fourth reminder.[15] Note that the lottery was randomized among those students who were still enrolled at the university (sample size $n = 472$), corresponding to 83% of participants in the baseline survey.[16]

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, including a multitude of courses and team sports. 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) poor, is taken from the follow-up survey at the end of the second semester.[17]

6.2 | Descriptive statistics

Table 5 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 5 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. Also note that the differential response across treated and non-treated compliers corresponds to the mean difference in response divided by the mean difference in the treatment, which amounts to 21%. 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 is 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').[18] The first stage effect is 0.084,

[15] The lottery emails were sent directly after the respective survey emails in order to avoid an additional reminder effect of the lottery.

[16] The overall costs of such a strategy for tackling survey attrition depend on the magnitude of the cash incentives and the response rate. In our cash lottery, with incentives being either 0 CHF (20% chance) or ranging from 10 to 200 CHF with equal probability and 25% chance of winning, the maximum average cost per student is CHF 21.03 if all students with a positive lottery value respond. De facto, the average cost amounted to CHF 12.57 in our experiment. This instrument is arguably more expensive than potential alternatives, like the number of randomly assigned reminder emails or phone calls. A mixture of cash- and contact-based instruments may at the same time satisfy the support requirements of the instrument and budget constraints for the generation of the instruments.

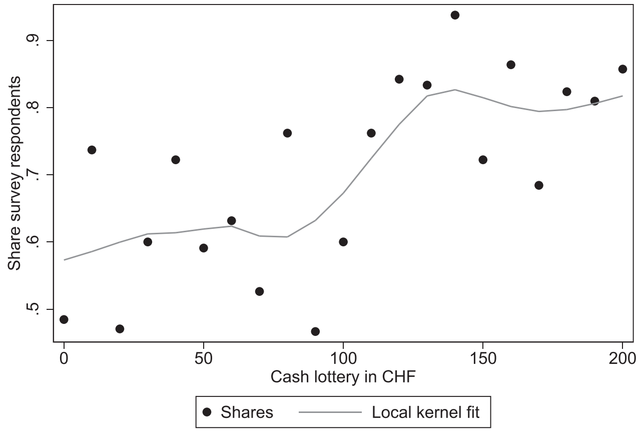
[17] For only two out of all respondents of the follow-up survey, self-reported health is missing (item non-response). The two missing values were set to the mean of self-assessed health among the 470 survey respondents without item non-response. Alternatively, deleting these two observations led to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

[18] The nonparametric specification test of (Sant' Anna & Song, 2019) does not reject our propensity score model for the first instrument at conventional levels of significance. The test yields p-values of roughly 40% or more when varying the choice of tuning parameters (see the 'pstest' package for R).

TABLE 5 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)
Cash incentives to visit gym	–	–	–	–	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.



kernel = epanechnikov, degree = 0, bandwidth = 17.55

FIGURE 1 Follow up Survey Response

implying that the cash incentive instrument increases the probability to visit the gym at least once by 8.4 percentage points (which corresponds to the complier share),[19] with a bootstrap p-value of 0.06. Table 6 reports the LATE and ATE estimates based on (20) and (21), respectively, using the semiparametric approach ('semiparametric'), which performed better in simulations with a modest sample size than fully nonparametric estimation, see Section 5. The table provides the estimates when using weighting functions (18) and (19), see 'w1' and 'w2'. We also consider semiparametric LATE/ATE estimation requiring response to be MAR and thus only selective with respect to observables Z_1 and X (rather than unobservables), see the discussion in Section 5. For the LATE, the estimator ('MAR') corresponds to the sample analogue of (30). For the ATE, we first compute the conditional LATEs given X by estimating $E[Y|R=1,Z_1=1,X]-E[Y|R=1,Z_1=0,X]/E[D|Z_1=1,X]-E[D|Z_1=0,X]$ based on regression and then average over the distribution of X in the total population to obtain the

[19] This number differs slightly from that in Table 5 since we control for covariates X.

TABLE 6 Application

semiparametric LATE estimation						semiparametric ATE estimation					
effects	**est**	**pval**	**differences**	**est**	**pval**	**effects**	**est**	**pval**	**differences**	**est**	**pval**
IV w1	0.57	0.86	IV w1 – MAR	−0.45	0.51	IV w1	0.43	0.94	IV w1 – MAR	−0.91	0.61
IV w2	0.61	0.94	IV w2 – MAR	−0.41	0.47	IV w2	0.53	0.98	IV w2 – MAR	−0.81	0.67
MAR	1.01	0.46				MAR	1.34	0.47			

Note: P-values ('pval') are based on the quantiles of the bootstrapped effects using 1999 bootstrap replications. 'est' are the estimated effects or differences in effects. 'IV' refers to semiparametric estimation based on (20) and (21) for the LATE and ATE, respectively, with parametric first step estimators. 'w1' and 'w2' stands for weighting based on (18) and (19), respectively. 'MAR' refers to LATE/ATE estimation when assuming response to be selective with respect to observables only. The outcome 'self-reported health' ranges from 1: very good to 5: poor.

ATE. We also report the differences between our instrument-based methods for tackling attrition and the MAR approaches ('differences'). Besides the effects and their differences, Table 6 gives the bootstrap p-values based on the quantiles of the resampled distribution of the effect estimates (1999 replications), see equation (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, making t-statistics misleading.

The LATE MAR estimator is positive (1.01 points), implying a decrease in self-assessed health if taken at face value, but with a p-value of 0.46 (or 46%) far from being statistically significant at any conventional level. The LATE estimates based on our instrument-based attrition correction, amounting to 0.57 and 0.61, are insignificant, too.[20] So are the differences in the effects of our methods and LATE MAR. A similar pattern arises for the estimation of the ATE. While the absolute differences in the point estimates with instrument-based attrition correction and under MAR are somewhat larger then for the LATE, they are nevertheless statistically insignificant. This also applies to any of the ATE estimates. 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 unobservables. We acknowledge that this may be due to the low precision of our estimates rooted in the small sample size and complier share. This suspicion is corroborated by the fact that the effect estimates vary substantially across the 100 support points η of the conditional distribution function Ψ_d used for approximating the integral over η in Lemma 2: The standard deviations of the estimated LATE and ATE across η amount to 25.07 and 26.76, respectively.

7 | CONCLUSION

This paper developed a nonparametric approach for identifying average treatment effects in the presence of both treatment endogeneity and attrition/non-response, using a binary instrument for the binary treatment and an instrument with rich (in general continuous) support for attrition. Furthermore, we proposed non- and semiparametric estimators based on the sample analogs of our identification results and investigated their performance in a simulation study. As an empirical illustration, we estimated the effect of gym training on students' self-assessed health at a Swiss University, 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.

ACKNOWLEDGMENT

We acknowledge support by the German Science Foundation SFB 884 project B5. Hans Fricke worked on this project before joining Amazon.com Inc.

OPEN RESEARCH BADGES

This article has earned an Open Data Badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available at [http://qed.econ.queensu.ca/jae/datasets/fricke001/].

[20] Our point estimates lie within the lower and upper bounds (0.28 and 1.29, respectively) based on Proposition 1 of Chen and Flores (2015) for the set identification of the LATE on compliers that respond irrespective of their treatment. This approach assumes a valid instrument for the treatment and monotonicity of response in the treatment, but neither requires a second instrument for response nor additive separability of observed and unobserved terms in the outcome equation.

REFERENCES

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

Abowd, J. M., Crepon, B., & Kramarz, F. (2001). Moment estimation with attrition: An application to economic models. *Journal of the American Statistical Association*, *96*, 1223–1230.

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

Angrist, J., & Fernández-Val, I. (2010). *Extrapolate-ing: External validity and overidentification in the late framework. (NBER working paper 16566)*. Cambridge, MA: National Bureau of Economic Research.

Angrist, J., Imbens, G., & Rubin, D. (1996). Identification of causal effects using instrumental variables. *Journal of American Statistical Association*, *91*, 444–472. (with discussion).

Aronow, P. M., & Carnegie, A. (2013). Beyond late: Estimation of the average treatment effect with an instrumental variable. *Political Analysis*, *21*, 492–506.

Behaghel, L., Crépon, B., Gurgand, M., & Le Barbanchon, T. (2015). Please call again: Correcting nonresponse bias in treatment effect models. *Review of Economics and Statistics*, *97*, 1070–1080.

Blanco, G., Chen, X., Flores, C. A., & Flores-Lagunes, A. (2016). *Average and quantile effects of training on employment and unemployment spells: A bounds analysis in the presence of censoring and noncompliance. (Working Paper)*. Syracuse, NY: Syracuse University.

Castiglioni, L., Pforr, K., & Krieger, U. (2008). The effect of incentives on response rates and panel attrition: Results of a controlled experiment. *Survey Research Methods*, *2*, 151–158.

Charness, G., & Gneezy, U. (2009). Incentives to exercise. *Econometrica*, *77*(3), 909–931.

Chen, X., & Flores, C. A. (2015). Bounds on treatment effects in the presence of sample selection and noncompliance: The wage effects of job corps. *Journal of Business and Economic Statistics*, *33*, 523–540.

DiNardo, J., McCrary, J., & Sanbonmatsu, L. (2006). Constructive proposals for dealing with attrition: An empirical example, *Working Paper*.

Dong, Y. (2019). Regression discontinuity designs with sample selection. *Journal of Business and Economic Statistics*, *37*, 171–186.

Farrell, L., & Shields, M. A. (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., Gottschalk, P., & Moffitt, R. (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., & Rubin, D. (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., Lechner, M., & Steinmayr, A. (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., & Huber, M. (2014). Treatment evaluation with multiple outcome periods under endogeneity and attrition. *Journal of the American Statistical Association*, *109*, 1697–1711.

Frumento, P., Mealli, F., Pacini, B., & Rubin, D. B. (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. A., & Wise, D. A. (1979). Attrition bias in experimental and panel data: The gary income maintenance experiment. *Econometrica*, *47*(2), 455–473.

Hayfield, T., & Racine, J. (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. J. (1979). Sample selection bias as a specification error. *Econometrica*, *47*, 153–161.

Hirano, K., Imbens, G., & Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, *71*, 1161–1189.

Horvitz, D. G., & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, *47*, 663–685.

Hsu, J. W., Schmeiser, M. D., Haggerty, C., & Nelson, S. (2017). The effect of large monetary incentives on survey completion: Evidence from a randomized experiment with the survey of consumer finances. *Public Opinion Quarterly*, *81*, 736–747.

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., & Mellace, G. (2015). Testing instrument validity for late identification based on inequality moment constraints. *Review of Economics and Statistics*, *97*, 398–411.

Imai, K. (2008). Sharp bounds on the causal effects in randomized experiments with 'truncation-by-death'. *Statistics & Probability Letters*, *78*, 144–149.

Imbens, G., & Angrist, J. (1994). Identification and estimation of local average treatment effects. *Econometrica*, *62*, 467–475.

Jacob, B., McCall, B., & Stange, K. (2018). College as country club: Do colleges cater to students' preferences for consumption? *Journal of Labor Economics*, *36*, 309–348.

Janssen, I., & Leblanc, A. G. (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.

Lee, D. (2009). Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *Review of Economic Studies*, *76*, 1071–1102.

Lee, S., & Salanié, B. (2018). Identifying effects of multivalued treatments. *Econometrica*, *86*, 1939–1963.

Little, R., & Rubin, D. (1987). *Statistical analysis with missing data*. New York: Wiley.

MacKinnon, J. G. (2006). Bootstrap methods in econometrics. *The Economic Record*, *82*, S2–S18.

Machado, C., Shaikh, A. M., & Vytlacil, E. J. (2019). Instrumental variables and the sign of the average treatment effect. *Journal of Econometrics*, *212*, 522–555.

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

Mourifié, I., & Wan, Y. (2017). Testing local average treatment effect assumptions. *The Review of Economics and Statistics*, *99*, 305–313.

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

Pforr, K., Blohm, M., Blom, A. G., Erdel, B., Felderer, B., Frässdorf, M., ..., & Rammstedt, B. (2015). Are incentive effects on response rates and nonresponse bias in large-scale, face-to-face surveys generalizable to germany? evidence from ten experiments. *Public Opinion Quarterly*, *79*(3), 740–768.

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

Robins, J. M., Rotnitzky, A., & Zhao, L. P. (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.

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

Sant' Anna, P. H. C., & Song, X. (2019). Specification tests for the propensity score. *Journal of Econometrics*, *210*, 379–404.

Schneider, S., & Becker, S. (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.

Schochet, P. Z., Burghardt, J. A., & McConnell, S. M. (2000). *National job corps study: Methodological appendix on the short-term impact analysis. (Mathematica Policy Research, report 8140-520)*. Princeton, NJ.

Schwiebert, J. (2012). Semiparametric estimation of a sample selection model in the presence of endogeneity. unpublished manuscript.

Semykina, A., & Wooldridge, J. (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*. London: Chapman and Hall.

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., Leblanc, A. G., Carson, V., Connor Gorber, S., Dillman, C., Janssen, I., ..., & Tremblay, M. S. (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., Rubin, D., & Mealli, F. (2009). Likelihood-based analysis of causal effects of job-training programs using principal stratification. *Journal of the American Statistical Association*, *104*, 166–176.

How to cite this article: Fricke H, Frolich M, Huber M, Lechner M. Endogeneity and non-response bias in treatment evaluation – nonparametric identification of causal effects by instruments. *J Appl Econ*. 2020;35: 481–504. https://doi.org/10.1002/jae.2764

APPENDIX A: PROOFS OF THEOREMS

A.1 | Preliminaries

We will repeatedly make use of

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

The proof is immediate via partitioning by types, as

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

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

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

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

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

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

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

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

$$
\begin{aligned}
&E\left[YRD\frac{\Pr(Z_1=1|X=x,\Psi_1=\bar{\eta})}{\Pr(Z_1=1|Z_2,X=x,\Psi_1=\bar{\eta})}\middle|X=x,\Psi_1=\bar{\eta},Z_1=1\right] \\
&\quad=\int E\left[YRD\frac{\Pr(Z_1=1|X=x,\Psi_1=\bar{\eta})}{\Pr(Z_1=1|Z_2,X=x,\Psi_1=\bar{\eta})}\middle|Z_2,X=x,\Psi_1=\bar{\eta},Z_1=1\right]dF_{Z_2|X=x,\Psi_1=\bar{\eta},Z_1=1} \\
&\quad=\int E\left[YRD|Z_2,X=x,\Psi_1=\bar{\eta},Z_1=1\right]dF_{Z_2|X=x,\Psi_1=\bar{\eta}} \\
&\quad=\int E\left[YRD|Z_2,X=x,Z_1=1\right]dF_{Z_2|X=x,\Psi_1=\bar{\eta}},
\end{aligned}
\tag{A.2}
$$

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:

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

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

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

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

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

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

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

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

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

By analogous derivations,

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

Analogously, we can derive

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

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

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

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

Similarly, we can derive

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

Now putting all results together, we obtain

$$\frac{1}{\bar{\eta}}\frac{E\left[\frac{YRD}{E\left[Z_1|Z_2,X=x,\Psi_1=\bar{\eta}\right]}\frac{Z_1-E\left[Z_1|Z_2,X=x,\Psi_1=\bar{\eta}\right]}{1-E\left[Z_1|Z_2,X=x,\Psi_1=\bar{\eta}\right]}\middle| X=x, \Psi_1=\bar{\eta}\right]}{E\left[\frac{D}{E\left[Z_1|Z_2,X=x,\Psi_1=\bar{\eta}\right]}\frac{Z_1-E\left[Z_1|Z_2,X=x,\Psi_1=\bar{\eta}\right]}{1-E\left[Z_1|Z_2,X=x,\Psi_1=\bar{\eta}\right]}\middle| X=x, \Psi_1=\bar{\eta}\right]}$$

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

$$= \frac{1}{\bar{\eta}}\left\{\phi(1,x)\cdot\bar{\eta} + \int E\left[U\cdot 1\left(V \le F^{-1}_{V|X=x,T=c}(\bar{\eta})\right)\middle| X=x, T=c\right]\right\}$$

$$- \frac{1}{\bar{\eta}}\left\{\phi(0,x)\cdot\bar{\eta} + \int E\left[U\cdot 1\left(V \le F^{-1}_{V|X=x,T=c}(\bar{\eta})\right)\middle| X=x, T=c\right]\right\}$$

$$= \phi(1,x) - \phi(0,x) = E\left[Y^1 - Y^0|X=x, T=c\right] = E\left[Y^1 - Y^0|X=x\right].$$

Define

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

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

$$E\left[Y^1 - Y^0\right] = \int \left(\int_0^1 \frac{1}{\eta} \left(\Xi_1(X,\eta) - \Xi_0(X,\eta)\right) \frac{w(\eta,X)}{\int w(\eta,X)d\eta} \right) d\eta dF_X \tag{A.10}$$

and the local average treatment effect on the compliers as

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

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

$$E\left[Y^1 - Y^0 | T = c\right] = \frac{1}{E\left[\frac{D}{P}\frac{Z_1 - P}{1-P}\right]} \int \left(\int_0^1 \frac{1}{\eta} \left(\Xi_1(X,\eta) - \Xi_0(X,\eta)\right) \frac{w(\eta,X)}{\int w(\eta,X)d\eta} \right) E\left[\frac{D}{P}\frac{Z_1 - P}{1-P}\middle| X\right] d\eta dF_X. \tag{A.11}$$

More generally, replacing $E\left[\frac{D}{P}\frac{Z_1-P}{1-P}|X\right]$ in (A.11) by some function $g(X)$ of covariates X and $E\left[\frac{D}{P}\frac{Z_1-P}{1-P}\right]$ by $E[g(X)]$, respectively, identifies a weighted ATE for a specific target population, assuming that $|g(X)|$ is bounded from above and $E[g(X)] > 0$. Setting for instance $g(X) = \Pr(D = 1|X)$ and $E[g(X)] = \Pr(D = 1)$ identifies the average treatment effect on the treated (ATET),

$$\begin{aligned}
E\left[Y^1 - Y^0 | D = 1\right] &= \int E\left[Y^1 - Y^0 | X, T = c\right] dF_{X|D=1} \\
&= \int E\left[Y^1 - Y^0 | X, T = c\right] \frac{\Pr(D = 1|X)\, dF_X}{\Pr(D = 1)} \\
&= \frac{1}{\Pr(D = 1)} \int \left(\int_0^1 \frac{1}{\eta} \left(\Xi_1(X,\eta) - \Xi_0(X,\eta)\right) \frac{w(\eta,X)}{\int w(\eta,X)d\eta} \right) \Pr(D = 1|X)\, d\eta dF_X.
\end{aligned} \tag{A.12}$$

Setting $g(x) = E[g(X)] = 1$, on the other hand, yields the ATE on the total population and coincides with (A.10).

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

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

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 $f(\Psi_1|X)$, and the fourth line that for the estimation of $f(\Psi_0|X)$.

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.[21]

Calculations with first term only. The first term of the influence function is

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

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

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

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)\,(1-\pi(X))} \cdot \left\{ D\frac{w(\Psi_1, X)}{\Psi_1 \cdot f\,(\Psi_1|X)} + (1-D)\frac{w(\Psi_0, X)}{\Psi_0 \cdot f\,(\Psi_0|X)} \right\}\right)^2\right] \\
&= E\left[Y^2R^2 \cdot \left(\frac{Z_1 - \pi(X)}{\pi(X)\,(1-\pi(X))}\right)^2 \cdot \left\{ D\frac{w(\Psi_1, X)}{\Psi_1 \cdot f\,(\Psi_1|X)} + (1-D)\frac{w(\Psi_0, X)}{\Psi_0 \cdot f\,(\Psi_0|X)} \right\}^2\right] \\
&= E\left[Y^2R \cdot \left(\frac{Z_1}{\pi(X)^2} + \frac{1-Z_1}{(1-\pi(X))^2}\right) \cdot \left\{ D\frac{w^2(\Psi_1, X)}{\Psi_1^2 \cdot f^2\,(\Psi_1|X)} + (1-D)\frac{w^2(\Psi_0, X)}{\Psi_0^2 \cdot f^2\,(\Psi_0|X)} \right\}\right] \\
&= E\left[Y^2R \cdot \left(\frac{Z_1}{\pi(X)^2} + \frac{1-Z_1}{(1-\pi(X))^2}\right) \cdot D\frac{w^2(\Psi_1, X)}{\Psi_1^2 \cdot f^2\,(\Psi_1|X)}\right] \\
&\quad + E\left[Y^2R \cdot \left(\frac{Z_1}{\pi(X)^2} + \frac{1-Z_1}{(1-\pi(X))^2}\right) \cdot (1-D)\frac{w^2(\Psi_0, X)}{\Psi_0^2 \cdot f^2\,(\Psi_0|X)}\right] \\
&= \int E\left[Y^2RD\left(\frac{Z_1}{\pi(X)^2} + \frac{1-Z_1}{(1-\pi(X))^2}\right)\frac{w^2(\Psi_1, X)}{\Psi_1^2 \cdot f^2\,(\Psi_1|X)}\middle|\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}{(1-\pi(X))^2}\right)\frac{w^2(\Psi_0, X)}{\Psi_0^2 \cdot f^2\,(\Psi_0|X)}\middle|\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}{(1-\pi(X))^2}\right)|\Psi_1, X\right]\frac{w^2(\Psi_1, X)}{\Psi_1^2 \cdot 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}{(1-\pi(X))^2}\right)\middle|\Psi_0, X\right]\frac{w^2(\Psi_0, X)}{\Psi_0^2 \cdot f\,(\Psi_0|X)} \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}\,(\eta|X)} + \frac{\lambda_0(\eta, X)}{f_{\Psi_0|X}\,(\eta|X)}\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 $\lambda_0(\eta, X) = E[Y^2R(1-D)\left(\frac{Z_1}{\pi(X)^2} + \frac{1-Z_1}{(1-\pi(X))^2}\right)|\Psi_0 = \eta, X]$.

[21] 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.