Universität St.Gallen

Identification of average treatment effects in social experiments under different forms of attrition

Martin Huber

June 2010 Discussion Paper no. 2010-22

Department of Economics University of St. Gallen

Editor: Martina Flockerzi
University of St. Gallen
Department of Economics
Varnbüelstrasse 19
CH-9000 St. Gallen
Phone +41 71 224 23 25
Fax +41 71 224 31 35
Email vwaabtass@unisg.ch

Publisher: Department of Economics
University of St. Gallen
Varnbüelstrasse 19
CH-9000 St. Gallen
Phone +41 71 224 23 25
Fax +41 71 224 31 35

Electronic Publication: http://www.vwa.unisg.ch

Identification of average treatment effects in social experiments under different forms of attrition[1]

Martin Huber

Author's address:

Martin Huber, Ph.D.
SEW
Varnbüelstrasse 14
9000 St Gallen
Phone +41 71 2299
Fax +41 71 2302
Email Martin.Huber@unisg.ch
Website www.sew.unisg.ch

[1] I have benefited from comments by Eva Deuchert, Bernd Fitzenberger, Michael Lechner, Blaise Melly, and conference/seminar participants in Freiburg i. B. (research seminar), London (EALE 2010), and Fribourg (SSES 2010).

Abstract

As any empirical method used for causal analysis, social experiments are prone to attrition which may flaw the validity of the results. This paper considers the problem of partially missing outcomes in experiments. Firstly, it systematically reveals under which forms of attrition - in terms of its relation to observable and/or unobservable factors – experiments do (not) yield causal parameters. Secondly, it shows how the various forms of attrition can be controlled for by different methods of inverse probability weighting (IPW) that are tailored to the specific missing data problem at hand. In particular, it discusses IPW methods that incorporate instrumental variables when attrition is related to unobservables, which has been widely ignored in the experimental literature.

Keywords

Experiments, attrition, inverse probability weighting.

JEL Classification

C21, C93

1 Introduction

Causal inference based on experiments, which dates at least back to Neyman (1923) and Fisher (1925, 1935), is a cornerstone of the evaluation of policy interventions. It has been used in many different fields of research such as medicine, welfare policies, labor economics, eduction, and development economics, see for instance the literature surveys in Duflo (2006), Harrison and List (2004), and Imbens and Wooldridge (2009). If well conducted and appropriate to the research question, experiments are widely regarded to be the most reliable source of causal inference, see for instance Cochran and Chambers (1965), Freedman (2006), Rubin (2008), and Imbens (2009), as they invoke a minimum of identifying assumptions. They neither impose functional form assumptions as regression models nor particular correlations between observables and unobservables which have to be assumed in observational studies. However, as any empirical method, experiments are prone to attrition which may flaw the validity of the results, see the discussion in Hausman and Wise (1979).

In this paper, we consider the problem that the outcome of interest is only partially observed due to attrition, whereas the treatment and several socio-economic characteristics, which are typically measured in baseline surveys prior to the intervention, are fully observed. The missing outcome problem arises for instance when individuals with known pre-treatment characteristics are randomly assigned to an active labor market policy (such as a training), but some of them do not participate in a follow-up survey that measures labor market success (e.g., employment or income) several months or years later due to reluctance or relocation. Similar problems are inherent in clinical trials when some of the participants randomly assigned to medical treatments pass away ('truncation by death') before the health outcome is measured. Finally, suppose that high-school students are randomly provided with private school vouchers and that we are interested in their scores obtained in college entrance examinations several years later. Attrition in the outcome arises if a subsample of students decides not to take the exam.

Various remedies have been proposed to deal with attrition in outcome data. Multiple im-

putation of missing values goes back to Rubin (1977, 1978), see also Rubin (1996) for a more recent review. Based on Bayesian techniques, the idea is to use multiple attrition models to impute multiple sets of plausible values for the missing data. This allows computing a probability interval for the parameter of interest. Several studies use single imputation methods such as regression adjustments to correct for attrition. E.g., Hausman and Wise (1979) use a probability model of attrition in conjunction with a random effects model of individual response in their evaluation of the Gary Income Maintenance experiment. Angrist, Bettinger, and Kremer (2006) analyze the effects of school vouchers on college entrance examinations in Columbia and apply tobit regression to control for the fact that voucher winners are more likely to take the exams than voucher losers. Another approach is based on weighting observations according to their likelihood to respond, i.e., by the inverse of their conditional response probability, see for instance Scharfstein, Rotnitzky, and Robins (1999). The idea of inverse probability weighting (IPW) goes back to Horvitz and Thompson (1952), who first proposed an estimator of the population mean in the presence of non-randomly missing data.

Barnard, Frangakis, Hill, and Rubin (2003) use a principal stratification framework (see Frangakis and Rubin, 2002) to estimate treatment effects under attrition (and further missing data and non-compliance problems) by means of a parametric mixture model. Still based on principal stratification, Mealli and Pacini (2008) exploit discrete instruments to identify effects for particular subgroups under various assumptions. Finally, the estimation of nonparametric bounds (see Horowitz and Manski, 1998, 2000) does not require a model for attrition at the cost of sacrificing point identification even for particular subpopulations. Bounds on experimental treatment effects have been estimated by Angrist, Bettinger, and Kremer (2006), Lee (2009), who evaluates the effects of the Job Corps program on potential wages but observes the latter only conditional on employment, and Grogger (2009), who assesses the effectiveness Connecticuts Jobs First experiment and faces attrition due to relocation to a different state. See also Zhang and Rubin (2003) and Zhang, Rubin, and Mealli (2008) who discuss the identification of bounds in a principal stratification framework.

This paper makes two contributions to the literature on attrition in social experiments. Firstly, it reveals systematically under which forms of attrition - in terms of its relation to observable or unobservable factors - experiments do (not) yield causal parameters, as a comprehensive discussion on attrition in experiments and its implications for identification is still lacking. Starting from a general treatment effect model, it makes explicit and formally discusses under which conditions experiments identify average treatment effects on the entire population and/or on the subpopulation of respondents.

Secondly, the paper shows how the different forms of attrition can be controlled for by IPW methods, i.e., by reweighing observations by the inverse of their conditional response and/or treatment probabilities. Depending on whether attrition is related to all or subsets of observable and unobservable variables, we will apply different weighting approaches, each of which is tailored to the specific attrition problem at hand. This provides practitioners with straightforward solutions depending on the suspected missing data problem.

The use of IPW to control for attrition related to observables, i.e., when outcomes are missing at random (MAR, see Rubin 1976), is well established in the literature, see for instance Robins and Rotnitzky (1995), Robins, Rotnitzky, and Zhao (1995), Rotnitzky and Robins (1995), Scharfstein, Rotnitzky, and Robins (1999), and Wooldridge (2002, 2007). In contrast, the case when attrition is related to unobservables such that identification requires an instrument for attrition (which does not directly affect the outcome variable) has been widely ignored in experiments. One of the very rare examples is DiNardo, McCrary, and Sanbonmatsu (2006) who use conventional sample selection correction techniques based on regression, see Heckman (1976). The present work is the first that discusses the usefullness and application of IPW under attrition on unobservables in an experimental context. This approach is closely related to Huber (2009), who uses IPW to control for sample selection and attrition in observational studies.

The remainder of this paper is organized as follows. The next section introduces a general treatment effect model along with attrition. Section 3 discusses identification under random attrition and attrition related to observables. Identification under attrition related to unobserv-

ables is treated in Section 4. Section 5 presents simulation studies based on both generated and empirical data. An application to a randomized introduction of a new health policy in Mexico is provided in Section 6. Section 7 concludes.

2 Model

Let D denote a treatment indicator, either 1 (treatment) or 0 (non-treatment), which is randomly assigned to an i.i.d. sample of n units, indexed by $i = 1, ..., n$. We are interested in the effect of D on some outcome variable Y. Utilizing the potential outcome framework of Rubin (1974), we denote the potential outcome for individual i and some hypothetical treatment $D = d$ as Y_i^d, where $d \in \{0, 1\}$. The difference $Y_i^1 - Y_i^0$ would identify the individual treatment effect, but is unknown to the researcher, because each individual is either treated or not treated and cannot appear in both states of the world at the same time.

However, under particular assumptions a randomized experiment allows identifying treatment effects by the fact that the potential outcomes are independent of the treatment assignment. Throughout this paper we will therefore rule out any interaction effects between the individuals participating in the experiment such as spill over, peer, or general equilibrium effects. This implies the validity of the Stable Unit Treatment Valuation Assumption (SUTVA), see for instance Rubin (1990). Furthermore, we will assume that that treatment compliance is perfect, i.e., everybody being assigned takes the treatment, everybody not assigned does not. Even though we are fully aware that interaction effects and noncompliance in experiments (see for instance Robins and Tsiatis, 1991) do occur in practice, they are beyond the scope of this paper. In the subsequent discussion we will exclusively focus on the identification problems related to attrition in the outcome variable.

Under random treatment assignment the expected potential outcomes are equal to the expected conditional outcomes given the treatment. Formally, $E[Y^d] = E[Y|D = d]$ for $d \in \{0, 1\}$. Therefore, the average treatment effect (ATE), denoted as Δ, is identified by

$E[Y|D = 1] - E[Y|D = 0]$ and is consistently estimated by the mean difference of treated and nontreated outcomes in the sample. Causal inference becomes less straightforward when the outcome variable is only partially observed due to attrition. The problems arising for identification and the remedies that may be used depend on how attrition is related to the treatment and the other parameters affecting the outcome.

To formally discuss the various forms of attrition, we consider a general treatment effect model. Assume that the outcome Y is an unknown function of the treatment, a vector of observed covariates X, and an unobserved term U.

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

where $\varphi(\cdot)$ is a general function. Throughout this paper we will maintain that the treatment is randomly assigned such that $(X, U) \perp D$, where '$\perp$' denotes independence. When using the potential outcomes notation, this is implies that $Y^1, Y^0 \perp D$. In Section 3, we will also assume that X contains at least one continuously distributed variable. In Section 4, X may or may not be continuous.

This model provides us with a useful framework for the evaluation of policy interventions. Consider for instance the identification of the effect of vouchers for private schooling (D) to which high school students are randomly assigned on test scores in college entrance examinations (Y) several years later. Empirical evidence suggests that private schooling has an effect on test scores, see Angrist, Bettinger, and Kremer (2006). Thus, we suspect the test scores to be a function of the treatment, but also of observed baseline characteristics (X) such as age and gender, which are usually provided in surveys accompanying randomized trials. Furthermore, also unobserved factors U such as motivation most likely influence the test scores. As a second example, consider labor market experiments where individuals are randomly assigned into a training, see for instance Bloom, Orr, Bell, Cave, Doolittle, Lin, and Bos (1997). The labor market outcomes (Y), e.g., employment, unemployment, or income, are a function of training

(D), socio-economic characteristics like age, education, and gender (X), and unobservables (U) such as innate ability.

To introduce attrition into our framework, let R denote a binary response variable which is 1 if Y is observed (non-attrition) and 0 otherwise. In contrast, we will assume that (X, D) is observed for all individuals. The fact that only $Y|R = 1$ is known instead of Y may flaw the validity of experimental results. The experiment bears external validity if it identifies the ATE on the entire population ($\Delta = E[Y^1] - E[Y^0]$) in spite of attrition. It bears internal validity if the ATE on the respondents, $\Delta_{R=1} = E[Y^1|R = 1] - E[Y^0|R = 1]$, is identified. Whether external and/or internal validity holds depends on the nature of attrition. The following two sections will impose different assumptions on the relation between attrition and observed and unobserved factors in the treatment effect model and will discuss the implications for identification. When identification fails, we will propose IPW methods to correct for attrition bias and also discuss the required conditions.

3 Identification under random attrition and attrition related to observables

The most innocuous form of attrition is the case when outcomes are missing completely at random (MCAR), see for instance Rubin (1976) and Heitjan and Basu (1996). MCAR says that attrition is not related with any observed or unobserved parameter in the treatment effect model. After considering this benchmark case we will systematically investigate more severe attrition problems.

Assumption 1: R is random.

$R \perp (D, X, U)$.

Assumption 1 states that attrition is independent of both observed and unobserved variables. This implies that the potential outcomes are independent of the response mechanism, $Y^1, Y^0 \perp R$,

and that the potential outcomes and the response mechanism are jointly independent of the treatment assignment, $(Y^1, Y^0, R) \perp D$. To see the implications for identification, note that the potential outcome under treatment $D = d$ for individual i is $Y_i^d \equiv \varphi(d, X_i, U_i)$ for $d \in \{0, 1\}$. Furthermore, let F_A denote the cdf of a random variable A and $F_{A|B}$ the conditional cdf given B. By MCAR,

$$\int \varphi(d, X, U) dF_{U,X|D=d,R=1} = \int \varphi(d, X, U) dF_{U,X|D=d} = \int \varphi(d, X, U) dF_{U,X},$$

where the first equality follows from Assumption 1 and the second from random assignment. Therefore, the experiment bears external validity and identifies the ATE in spite of attrition:

$$\begin{aligned} & \int \varphi(1, X, U) dF_{U,X|D=1,R=1} - \int \varphi(0, X, U) dF_{U,X|D=0,R=1} \\ = \; & E[Y|D = 1, R = 1] - E[Y|D = 0, R = 1] = E[Y|D = 1] - E[Y|D = 0] \\ = \; & E[Y^1] - E[Y^0] = \Delta. \end{aligned}$$

As a first deviation from MCAR, we will now assume that response is a function of the treatment, but not of any other parameter in the treatment effect model.

Assumption 2: R is a function of D and a random component V.
(2a) $R = I\{\zeta(D, V) \geq 0\}$,
(2b) $V \perp (X, U)$.

$I\{\cdot\}$ denotes the indicator function and $\zeta(\cdot)$ is a general function. We assume V to be an unobserved term that is independent of (X, U), see (2b). By (2a), D shifts R such that $\Pr(D = 1|R = 1) \neq Pr(D = 1|R = 0)$. As D also affects Y, it follows that $E[Y|R = 1] \neq E[Y]$ because the share of treated individuals changes due to attrition. However, this does not affect identification, because the distribution of (X, U) is not related to the response behavior. This implies that $(Y^1, Y^0, R^1, R^0) \perp D$, where R^d denotes the hypothetical response for $D = d$, and $Y^1, Y^0 \perp R|D$.

As under Assumption 1, it holds that

$$\int \varphi(d,X,U)dF_{U,X|D=d,R=1} = \int \varphi(d,X,U)dF_{U,X|D=d} = \int \varphi(d,X,U)dF_{U,X}.$$

where the first equality follows from Assumption 2 and the second from random assignment. The experiment is again externally valid and identifies the ATE.

The forms of attrition considered under Assumptions 1 and 2 are unlikely to hold in many, if not most social experiments. Empirical evidence suggests that response behavior is often related to the treatment *and* other observed characteristics, see for instance Hausman and Wise (1979), Fitzgerald, Gottschalk, and Moffitt (1998), and Grilo, Money, Barlow, Goddard, Gorman, Hofmann, Papp, Shear, and Woods (1998). These characteristics X are typically measured in baseline surveys prior to the intervention and commonly include gender, age, education, and other socio-economic variables.

In the remainder of this section, we will assume that response is a function of the treatment and the covariates. In a first step, we impose a very particular relationship between X and D, namely independence conditional on response. This case is primarily chosen for illustrative reasons rather than practical relevance. Interestingly, it entails internal validity of the experiment while external validity does no longer hold without controlling for attrition.

Assumption 3: R is a function of D, X, and V.
(3a) $R = I\{\zeta(D,X,V) \geq 0\}$,
(3b) $V \perp (X,U)$,
(3c) $X \perp D|R = 1$.

By Assumption 3, attrition affects the distributions of D and X, which are, however, not related to each other even conditional on response, see (3c). This implies that the distributional change in X is equal across treatment states. As U does not affect the response the joint distribution of (X,U) conditional on $R = 1$ is independent of D. Thus, $\int \varphi(d,X,U)dF_{X,U|D=d,R=1} = \int \varphi(d,X,U)dF_{X,U|R=1}$. The mean potential outcome of respondents is equal to the average con-

ditional outcome given $D = d$ among respondents. Therefore, the experiment identifies the ATE on respondents ($\Delta_{R=1}$) and is internally valid:

$$\begin{aligned} & \int \varphi(1, X, U) dF_{U,X|D=1,R=1} - \int \varphi(0, X, U) dF_{U,X|D=0,R=1} \\ = \quad & E[Y|D = 1, R = 1] - E[Y|D = 0, R = 1] = E[Y^1|R = 1] - E[Y^0|R = 1] = \Delta_{R=1}. \end{aligned}$$

However, it is not externally valid, which would require that $\Delta_{R=1} = \Delta$. The latter does not hold because the distribution of X is not the same for respondents and non-respondents such that $\int \varphi(d, X, U) dF_{X,U|R=1} \neq \int \varphi(d, X, U) dF_{X,U}$.

Under certain conditions, the ATE on the entire population is identified by weighting respondents according to the likelihood that their observed characteristics appear in the entire population. To this end, we define the response propensity score (see Rosenbaum and Rubin, 1983), i.e., the conditional response probability given (D, X), as $p(D, X) \equiv \Pr(R = 1|D, X)$. Furthermore, we impose the following common support restriction:

Assumption 3': Positive response propensity score.
$\Pr(R = 1|D = d, X = x) > c$ for all $x \in \mathcal{X}$, $d \in \{0, 1\}$, $c > 0$.

$\mathcal{X}$ denotes the support of X. Assumption 3' states that for any (X, D), the response probability must be bounded away from zero, otherwise outcomes are never observed for particular combinations of the treatment and the covariates. This allows us to reestablish external validity of the experiment by IPW as suggested in Proposition 1.

Proposition 1

Under Assumptions 3 and 3', the ATE is identified by

$$\Delta \quad = \quad E\left[\frac{R \cdot Y}{p(1, X)}|D = 1\right] - E\left[\frac{R \cdot Y}{p(0, X)}|D = 0\right] \tag{2}$$

Proof: See Appendix A.1.

Thus, weighting observations by the inverse of their respective response propensity score identifies the ATE. The idea of using IPW to control for attrition or similar selection problems goes back to Horvitz and Thompson (1952), who proposed an estimator of the population mean when data are missing non-randomly. IPW has been frequently applied when the attrition process is assumed to depend only on observables, i.e., when outcomes are missing at random (MAR) in the notation of Rubin (1976). Formally, the MAR requires that $\Pr(R = 1|D, X, Y) = \Pr(R = 1|D, X)$, or equivalently, that $Y \perp R|D, X$. E.g., Robins and Rotnitzky (1995), Robins, Rotnitzky, and Zhao (1995), Rotnitzky and Robins (1995), and Scharfstein, Rotnitzky, and Robins (1999) use IPW to adjust for missing data in regression models. Wooldridge (2002) considers IPW M-estimation of missing data models and Proposition 1 fits into his general framework as a special case.

To make our framework more general, we relax Assumption 3 somewhat by omitting Assumption (3c). This allows the gradient of X on the response to differ across treatment states. E.g., one could imagine that in a school voucher experiment, private schools ($D = 1$) are equally successful in sending younger and older students (X=age) to college entrance examinations ($R = 1$), whereas public schools ($D = 0$) more likely send older students. This would change the distribution of X across treatments among test takers.

Assumption 4: R is a function of D, X, and V.
(4a) $R = I\{\zeta(D, X, V) \geq 0\}$,
(4b) $V \perp (X, U)$

Without further assumptions, $\int \varphi(d, X, U) dF_{X,U|D=d,R=1} \neq \int \varphi(d, X, U) dF_{X,U|R=1}$. Thus, internal validity does no longer hold because the distribution of X generally differs across treatment states among respondents such that the effect of D is confounded. It, however, still holds that

$$\int \varphi(d, X, U) dF_{U|D=d,X=x,R=1} = \int \varphi(d, X, U) dF_{U|D=d,X=x}$$
$$= \int \varphi(d, X, U) dF_{U|X=x} \quad \text{for all } x \in \mathcal{X},$$

where the first equality follows from the randomness of response conditional on (X, D) implied

by Assumption 4, which satisfies MAR, and the second from the random assignment. Note that this would also hold if we relaxed (4b) somewhat to $V \perp U|X$.

It follows that the mean potential outcome among respondents is

$$\int \int \varphi(d, X, U) dF_{U|D=d,X=x,R=1} dF_{X|R=1}.$$

This allows us to identify the ATE on the respondents and to reestablish internal validity. Similarly to the response propensity score, we define the treatment propensity score among respondents, i.e., the conditional treatment probability given X, as $\pi(X) \equiv \Pr(D = 1|X, R = 1)$ and impose the following common support assumption:

Assumption 4': Common support in the treatment propensity scores.
$c < \Pr(D = 1|X = x) < 1 - c$ for all $x \in \mathcal{X}$, $d \in \{0, 1\}$, $c > 0$.

Assumption 4 states that the treatment propensity score is bounded away from 0 and 1, which rules out arbitrarily large weights in the subsequent proposition that reestablishes internal validity of the experiment.

Proposition 2

Under Assumptions 4 and 4', the ATE on respondents is identified by

$$\Delta_{R=1} = E\left[\frac{D \cdot Y}{\pi(X)} - \frac{(1 - D) \cdot Y}{1 - \pi(X)}|R = 1\right]. \tag{3}$$

Proof: See Appendix A.2.

By reweighing the outcomes of respondents by the inverse of the (non-)treatment propensity score, we control for differences in the distributions of X across treatment states to identify $\Delta_{R=1}$. This is analogous to the application of IPW in a 'selection on observables' or 'conditional independence' framework, see for instance Hirano, Imbens, and Ridder (2003), which also adjusts

for differences in X that, however, exist even prior to attrition.

Under somewhat stronger common support conditions we can also identify the ATE on the entire population and reestablish external validity. To this end, note that the mean potential in the entire population is

$$\int\int \varphi(d, X, U) dF_{U|D=d,X=x,R=1} dF_X.$$

I.e., integrating over the distribution of X in the entire population identifies the potential outcomes and the ATE. As for Proposition 1, this requires that the response probability is bounded away from zero for any (X, D).

Proposition 3

Under Assumptions 3', 4, and 4', the ATE is identified by

$$\Delta = E\left[\frac{R \cdot D \cdot Y}{p(D, X) \cdot \pi(X)} - \frac{R \cdot (1 - D) \cdot Y}{p(D, X) \cdot (1 - \pi(X))}\right]. \tag{4}$$

Proof: See Appendix A.3.

Δ is identified by using $\pi(X)$ to adjust for differences in the distributions of X across D among respondents and $p(D, X)$ to control for differences in (D, X) between respondents and non-respondents. Note that the strong Assumption (4b) may be replaced by the less severe restriction $V \perp U | D, X$, which might be considerably more plausible in empirical applications. Even then, response is random conditional on (D, X), MAR is satisfied, and Propositions 2 and 3 still apply.

4 Identification under attrition related to unobservables

In the last section we considered various forms of attrition related to observables. In our treatment effect model, MAR requires that U and V, the unobserved terms in the outcome and re-

sponse equations, are independent, at least conditional on observed characteristics. This assumption will be no longer maintained in this section. Instead, we will assume attrition on unobservables by allowing for a nonzero correlation between U and V even conditional on D, X. Analogous to sample selection models (see Heckman 1974, 1976, 1979) - at least when identification is nonparametric (e.g., Das, Newey, and Vella, 2003, Newey, 2007, and Huber, 2009) - point identification hinges on the availability of an instrument that affects response but has no direct effect on the outcome.

Reconsider for instance the identification of the effect of school vouchers assigned by a lottery on test scores in college entrance examinations. If only a subpopulation takes the test and the participation probability is a function of unobserved motivation that is also correlated with tests scores even conditional on the treatment and observed characteristics as age, identification requires an instrument that influences test participation but has no direct effect on the test scores.

Assumption 5: Attrition related to unobservables.
(5a) $R = I\{\zeta(D, X, Z) \geq V\}$,
(5b) $\text{Cov}(U, V) \neq 0$ (and $\text{Cov}(U, V) \neq 0|D, X$)

By Assumption 5, R is a function of one element that is excluded in φ, namely the instrument Z. Due to the non-zero covariance of V and U, the effect of D on Y among respondents is confounded even conditional on X. Identification requires Z to be a good predictor for R, to contain at least one continuous element, and not to have a direct effect on the outcome. The following restrictions guarantee the validity of the instrument.

Assumption 5': Instrument for the response probability.
(5'a) $\text{Cov}(Z, R|X, D) \neq 0$ and $Y \perp Z|D, X$,
(5'b) $\Pr(R = 1|D = d) > c$, $c > 0$, $d \in \{1, 0\}$,
(5'c) $(U, V) \perp (D, Z)$,

(5'd) $F_V(t)$, the cdf of V, is strictly monotonic in the argument t.

(5'a) states that Z shifts R but is not directly related with Y. (5'b) rules out that being treated or nontreated predicts attrition perfectly. To see the usefulness of this assumption, assume the opposite such that units with $D = 0$ do never respond independent of the values of (X, Z). Obviously, the treatment effect cannot be evaluated as no comparisons with $D = 0$ are available in the subpopulation of respondents.

By (5'c), we impose that D, Z are jointly independent of the unobservables (U, V). Independence between (U, V) and D is satisfied by the randomization of the treatment if V is not a post-treatment variable. Still, it needs to be plausibly argued that $(U, V) \perp Z$. E.g., let us assume that we want to investigate the effect of a randomized training on the potential wages (Y) of married women. If only those report their potential wages who actually work $(R = 1)$, we only observe $Y|R = 1$. Non-wife income appears to be a valid instrument for female labor force participation if it is not related to unobserved motivation or ability. However, Assumption (5'c) is violated if the distribution of motivation and ability (and thus, of potential wages) differs for women with different levels of non-wife income. This may be the case if more motivated and more able women tend to have more motivated and more able spouses with a higher wage potential.

As also argued by DiNardo, McCrary, and Sanbonmatsu (2006), the instrument should ideally be randomly assigned in a similar way as the treatment. This would plausibly justify Assumption (5'c). E.g., in a follow-up telephone survey, Z may be the number of phone calls per experimental unit which is randomized prior to the treatment assignment. A higher number of attempted calls should increase the response probability while being unrelated with other factors under random assignment. Also financial incentives are likely to affect response behavior, see Castiglioni, Pforr, and Krieger (2008). Note that (5'c) could be relaxed to $(U, V) \perp (D, Z)|X$, which might be more plausible in applications without randomized instruments.

Note that $\Pr(S = 1|D, X, Z) = \Pr(\zeta(D, X, Z) \geq V) = F_V(\zeta(D, X, Z))$. By the monotonicity assumption (5'd) the likelihood to respond increases monotonically in ζ. Monotonicity is implic-

itly assumed in any linear index restriction frequently used in the parametric sample selection literature, see Heckman (1974, 1976, 1979) . It allows us to back out the distribution of V by pinning down (D, X, Z). By comparing individuals with the same response propensity score, we control for V and thus, also for the dependence between V and U. I.e., by fixing V, we rule out confounding of the treatment effect due to attrition related to unobservables. The response propensity score serves as a control function where the exogenous variation comes from Z. Control functions have been applied in semi- and nonparametric sample selection models, e.g., Ahn and Powell (1993) and Das, Newey, and Vella (2003) as well as in nonparametric models with endogeneity, see, for example, Newey, Powell, and Vella (1999), Blundell and Powell (2003), and Imbens and Newey (2003).

However, conditioning on the response propensity score alone does not suffice for causal inference. A first reason is that similar to Assumption 4, response is a function of X and D. Therefore, random treatment assignment does not necessarily entail independence of D and X among respondents as the distribution of X might differ across treatment states conditional on $R = 1$. Secondly, this is even more likely to be the case conditional on the response propensity score. To see this, note that individuals in different treatment states D but equal values in X and Z must have distinct response propensity scores. I.e., $\Pr(R = 1|D = 1, X = x, Z = z) \neq \Pr(R = 1|D = 0, X = x, Z = z)$. As we need to compare treated and nontreated individuals with identical response propensity scores to control for the attrition bias, these individuals necessarily differ with respect to (X, Z). Despite the randomization of the treatment in the entire population, identification requires conditioning on both the response propensity score *and* the covariates among respondents which will be discussed more formally below.

For notational ease, let $W \equiv (D, X, Z)$ and the response propensity score $p(W) \equiv \Pr(R = 1|D, X, Z)$. If Assumption (5'c) holds, U and D are independent conditional on $p(W)$ and X, which can be shown analogously to the proof of Theorem 1 in Newey (2007). Let $a(U)$ denote

any bounded function of U. Note that $\{R=1\} = \{F_V^{-1}(p(W)) \geq V\}$. Then,

$$\begin{aligned}
E\left[a(U)|D,X,p(W),R=1\right] &= E\left[E\left[a(U)|V,D,X,Z\right]|D,X,p(W),F_V^{-1}(p(W)) \geq v\right] \\
&= E\left[E\left[a(U)|V,X\right]|D,X,p(W),F_V^{-1}(p(W)) \geq v\right] \\
&= E\left[E\left[a(U)|V,X\right]|X,p(W),F_V^{-1}(p(W)) \geq v\right] \\
&= E\left[E\left[a(U)|V,X,p(W)\right]|X,p(W),R=1\right] \\
&= E\left[a(U)|X,p(W),R=1\right],
\end{aligned}$$

where the first equality follows from iterated expectations, the second and third from (5'c), and the last from a backward application of the law of iterated expectations.

Thus, treatment effects are identified by conditioning on the response propensity score and the covariates. To see this, note that the conditional ATE given X and $p(W)$ conditional on response is defined as

$$\begin{aligned}
\Delta_{R=1}(x,p(w)) &= \int \varphi(1,x,u)dF_{u|X=x,p(W)=p(w),R=1} \\
&- \int \varphi(0,x,u)dF_{u|X=x,p(W)=p(w),R=1} \\
&= E[Y^1|X=x,p(W)=p(w),R=1] - E[Y^0|X=x,p(W)=p(w),R=1].
\end{aligned}$$

$E[Y^d|X=x,p(W)=p(w),R=1]$ is the expected potential outcome for a hypothetical treatment d given X and $p(W)$ among respondents. By the conditional independence of U and D given $p(W)$ and X, it holds that

$$\begin{aligned}
&E[Y^d|X=x,p(W)=p(w),R=1] = \int \varphi(d,x,u)dF_{u|X=x,p(W)=p(w),R=1} \\
&= \int \varphi(d,x,u)dF_{u|D=d,X=x,p(W)=p(w),R=1} \\
&= E[Y|D=d,X=x,p(W)=p(w),R=1].
\end{aligned}$$

Hence, the expected *potential* outcome is equal to the expected *conditional* outcome given $D=d$.

The ATE on respondents $\Delta_{R=1}$ is identified by the integration over the marginal distributions of X and $p(W)$ in the subpopulation with observed outcomes.

$$\begin{aligned}
&\int\int [E[Y|D=1, X=x, p(W)=p(w), R=1] \\
&-E[Y|D=0, X=x, p(W)=p(w), R=1]] dF_{x|p(W)=p(w),R=1} dF_{p(w)|R=1} \\
&= \int\int [E[Y^1|X=x, p(W)=p(w), R=1] \\
&-E[Y^0|X=x, p(W)=p(w), R=1]] dF_{x|p(W)=p(w),R=1} dF_{p(w)|R=1} \\
&= E[Y^1 - Y^0|R=1] = \Delta_{R=1}. \qquad (5)
\end{aligned}$$

Therefore, the identification of $\Delta_{R=1}$ hinges on the common support of the treatment in X and $p(W)$. Δ is not identified without further assumptions, see also Newey (2007), as Y is not even observed when $R = 0$. However, under the additional restrictions that the response propensity score is positive for any $(X, p(W))$ and that Y is homoscedastic conditional on (D, X) one can even identify the ATE on the entire population. To this end, we impose the following assumption:

Assumption 5": Common support and separability.

(5"a) $\Pr(R=1|D=d, X=x, Z=z) > c,\ c > 0$, for all $x \in \mathcal{X}$, for all $Z \in \mathcal{Z}$,

(5"b) $c < \Pr(D=1|X=x, p(W)=p(w)) < 1-c$, for all $x \in \mathcal{X}$, for all $p(w) \in \mathcal{P}$, $c > 0$,

(5"c) $Y = \varphi(D, X) + U$.

$\mathcal{Z}, \mathcal{P}$ denote the support regions of Z and $p(W)$. Note that (5"a) is stronger than (5'b). Effects on the entire population could not be identified if there existed individuals with a response propensity score equal to zero as this would rule out suitable comparisons in the subpopulation of respondents. (5"c) decreases the generality of our model due to separability of the observed and unobserved terms, see also Das, Newey, and Vella (2003), but ensures homoscedasticity of Y given (D, X). This is required for the identification of Δ, as outlined further below. Similar to the last section, let $\pi(X, p(W))$ denote the treatment propensity score, $\pi(X, p(W)) \equiv \Pr(D =$

$1|X, p(W), R = 1)$. Propositions 4 and 5 show identification of the ATEs on the respondents and on the entire population.

Proposition 4

Under Assumptions 5, 5', and (5"b), the ATE on the respondents is identified by

$$\Delta_{R=1} = E\left[\frac{D \cdot Y}{\pi(X, p(W))}|R = 1\right] - E\left[\frac{(1-D) \cdot Y}{1 - \pi(X, p(W))}|R = 1\right]. \tag{6}$$

Proof: See Appendix A.4.

By weighting observations by the inverse of the (non-)treatment propensity score, we adjust for differences in the distributions of X and $p(W)$ between treated and nontreated respondents. Proposition 4 is similar to Proposition 3, with the exception that we have to additionally condition on the response propensity score in the treatment propensity score to control for attrition on unobservables. It becomes immediately obvious that this approach is different to Mealli and Pacini (2008) who control for attrition in experiments by conditioning on a (discrete) instrument directly, rather than $p(W)$. Under certain conditions they identify treatment effects for particular subgroups. As we assume a continuous instrument, we can even identify the ATE on the entire population, given that there is common support in the propensity scores.

Proposition 5

Under Assumptions 5, 5', and 5", the ATE is identified by

$$\Delta \quad = \quad E\left[\frac{R \cdot D \cdot Y}{p(W) \cdot \pi(X, p(W))}\right] - E\left[\frac{R \cdot (1-D) \cdot Y}{p(W) \cdot (1 - \pi(X, p(W)))}\right]. \tag{7}$$

Proof: See Appendix A.5.

The ATE on the entire population is identified based on reweighing observations (in addition to the inverse treatment propensity score) by the inverse of the response propensity score, i.e., by using the relative likelihood of a particular triple (D, X, Z) to appear in the entire population, as weighting function.

This result may seem surprising given the fact that outcomes are only partially observed and observed outcomes do generally not allow inferring on the unobserved outcomes. I.e., $E[Y|D = d, X = x, p(W) = p(w), R = 1] \neq E[Y|D = d, X = x, p(W) = p(w), R = 0]$ due to different conditional distributions of the unobserved term U. Nevertheless, Assumptions (5') and (5") imply that $\Delta_{R=1}(x, p(w)) = \Delta_{R=0}(x, p(w))$. To see this, note that $F_{U|D=d,X=x,p(W)=p(w),R=r} = F_{U|X=x,p(W)=p(w),R=r}$ for $r \in \{0, 1\}$ by (5'c) such that

$$\begin{aligned}
\Delta_{R=1}(x, p(w)) &= \int [\varphi(1, x) + u] dF_{U|X=x,p(W)=p(w),R=1} \\
&- \int [\varphi(0, x) + u] dF_{U|X=x,p(W)=p(w),R=1}, \\
\Delta_{R=0}(x, p(w)) &= \int [\varphi(1, x) + u] dF_{U|X=x,p(W)=p(w),R=0} \\
&- \int [\varphi(0, x) + u] dF_{U|X=x,p(W)=p(w),R=0}.
\end{aligned}$$

$\Delta_{R=1}(x, p(w))$ and $\Delta_{R=0}(x, p(w))$ only differ with respect to the integrals over different conditional distributions of U given $R = 1$ and $R = 0$, which cancel out in the subtractions by (5"c). Thus, $\Delta_{R=1}(x, p(w)) = \Delta_{R=0}(x, p(w))$. Therefore, reweighing the conditional treatment effects of the respondents according to the distribution of (D, X, Z) in the entire population identifies Δ.

For completeness, we will briefly discuss identification under a particular deviation from the previous model, assuming that response is a function of D, Z, and V, but is not related with the covariates X.

Assumption 6: Attrition related to unobservables.
(6a) $R = I\{\zeta(D, Z) \geq V\}$,
(6b) $\text{Cov}(U, V) \neq 0$

Under this particular form of attrition the response behavior is merely a function of the treatment and unobservables, but unrelated to the observed covariates. Whether Assumption 6 is plausible

depends on the evaluation problem at hand and may even be tested (by testing the explanatory power of X on R). The response propensity score is now $p(D,Z) \equiv \Pr(R=1|D,Z)$. We impose the following instrumental variable and common support assumptions.

Assumption 6': Instrument for response probability.

(6'a) $\text{Cov}(Z,R|D) \neq 0$ and $Y \perp Z|D$,

(6'b) $\Pr(R=1|D=d) > c$, $c>0$, $d \in \{1,0\}$,

(6'c) $(X,U,V) \perp (D,Z)$,

(6'd) $F_V(t)$, the cdf of V, is strictly monotonic in the argument t.

Assumption 6": Common support in the response and treatment propensity scores.

(6"a) $\Pr(R=1|D=d,Z=z) > c,\ c>0$, for all $x \in \mathcal{X}$, for all $Z \in \mathcal{Z}$,

(6"b) $c < \Pr(D=1|p(D,Z)=p(d,z)) < 1-c$, for all $x \in \mathcal{X}$, for all $p(d,z) \in \mathcal{P}$, $c>0$,

(6"c) $Y = \varphi(D,X) + U$.

Assumption 6' and 6" are straightforward modifications of 5',5", with the exception of 6'c, which assumes independence of the (D,Z) also w.r.t. X. Again, the treatment is independent by randomization, whereas the independence of Z and X may or may not be plausible and might be tested. By Assumptions 6, 6'c, and random treatment assignment, it holds that X is independent of D conditional on $p(D,Z)$ and $R=1$, because $F_{X|D,p(D,Z),R=1} = F_{X|D} = F_X$. Therefore, treatment effects are identified conditional on $p(D,Z)$ or on the simplified treatment propensity score $\pi(p(D,Z)) \equiv \Pr(D=1|p(D,Z))$, respectively. Similar to Propositions 4 and 5, the ATEs on the respondents and the entire population can be expressed by the following equations:

Proposition 6

Under Assumptions 6, 6', and (6"b), the ATE on the respondents is identified by

$$\Delta_{R=1} = E\left[\frac{D \cdot Y}{\pi(p(D,Z))}|R=1\right] - E\left[\frac{(1-D)\cdot Y}{1-\pi(p(D,Z))}|R=1\right]. \tag{8}$$

Proof: See Appendix A.6.

Proposition 7

Under Assumptions 6, 6', and 6", the ATE is identified by

$$\Delta = E\left[\frac{R \cdot D \cdot Y}{p(D,Z) \cdot \pi(p(D,Z))}\right] - E\left[\frac{R \cdot (1-D) \cdot Y}{p(D,Z) \cdot (1-\pi(p(D,Z)))}\right]. \quad (9)$$

Proof: See Appendix A.7.

In the last two sections we have covered several forms of attrition and provided a guideline on which weighting methods are appropriate under specific assumptions. In particular, the use of IPW incorporating instrumental variables when attrition is on unobservables has been discussed, a case widely ignored in the experimental literature. Even though Sections 3 and 4 aim at covering cases that are relevant in many empirical applications, the discussion is not exhaustive, as one can think of many different ways of modeling response behavior in experiments.

We conclude this section by briefly reviewing a somewhat different identifying restriction recently proposed by Imai (2009), which he calls the 'non-ignorability assumption'. The latter implies conditional independence of the treatment and the response in an experiment:

$$\Pr(R=1|D=1,Y=y,X=x) = \Pr(R=1|D=0,Y=y,X=x),$$

or, equivalently,

$$R \perp D|Y,X.$$

I.e., conditional on the outcome and observed characteristics, response does not depend on the treatment, whereas R is allowed to be related with Y, e.g., through a correlation of U and V when considering the treatment effect model. As the author argues, this is plausible when response behavior is strongly driven by the outcome variable (e.g. voters may be more willing to participate in post-election surveys than non-voters), whereas the treatment represents only a

mild intervention that is unlikely to affect attrition (e.g., a psychological voting stimulus). Imai provides a method that identifies the ATE under the non-ignorability assumption and applies it to a German election experiment.

5 Simulation studies

In this section, we run a horse race between the experimental mean difference estimator not controlling for attrition and IPW estimators assuming that attrition is related to observables and unobservables as treated under Assumptions 4 and 5, respectively. For this reason, we conduct simulation studies based on both generated and empirical data. Starting with the former scenario, we consider the following data generating process (DGP):

$$\begin{aligned} Y_i &= \alpha_1 D_i + \alpha_2 X_i + \alpha_2 D_i X_i + U_i, \\ Y_i &\quad \text{is} \quad \text{observed if } R_i = 1, \\ R_i &= I\{\beta_1 D_i + \beta_2 X_i + \beta_3 Z_i + V_i > 0\}. \end{aligned}$$

Apart from the treatment D, the covariate X, and the unobserved term U, the outcome Y also depends on an interaction term of D and X, which introduces effect heterogeneity with respect to X. X and Z are uniformly distributed with support regions $[-1, 1]$ and $[-1, 2]$, respectively. D is Bernoulli and either 1 or 0 with equal probability. (U, V) are drawn from a multivariate standard normal distribution. Their covariance is set to zero in the case of attrition on observables and to 0.8 under attrition on unobservables. The coefficients in the outcome equation are set to $\alpha_1 = \alpha_2 = 1, \alpha_3 = 0.25$. Under attrition on observables $\beta_1 = \beta_2 = 1$ and $\beta_3 = 0$. Under attrition on unobservables, $\beta_1 = \beta_2 = 0.5$ and $\beta_3 = 1$, i.e., R is also a function of the instrument Z which is excluded from the outcome equation.

We use normalized versions (such that weights add up to unity, see Imbens, 2004, and Busso, DiNardo, and McCrary, 2009b) of the sample analogs of Propositions 4 and 5 as estimators of

the ATE on the entire population (denoted as $\hat{\Delta}$) and on the respondents ($\hat{\Delta}_{R=1}$). The response and treatment propensity scores $p(W), \pi(X, p(W))$ are specified as probit models. We run 1999 Monte Carlo simulations for two different sample sizes ($n = 500, 2000$) and compare the accuracy of IPW estimators to simply taking mean differences of treated and nontreated outcomes among respondents.

The following two tables report the results for untrimmed IPW estimators as hardly any propensity score estimate in any Monte Carlo replication is close to the boundaries of 0 and 1. Therefore, methods incorporating propensity score trimming (see for instance Busso, DiNardo, and McCrary, 2009a, and Crump, Hotz, Imbens, and Mitnik, 2009) yield virtually the same results and are omitted in the paper, but available upon request.

Table 1: Attrition on observables, simulated data

	n=500							
	$\hat{\Delta}$	bias	s.e.	MSE	$\hat{\Delta}_{R=1}$	bias	s.e.	MSE
IPW obs	1.001	0.001	0.180	0.032	0.998	-0.002	0.112	0.012
mean difference	0.883	-0.117	0.135	0.032	0.851	-0.149	0.130	0.039
true effect (normalized)	1.000				1.000			
	n=2000							
	$\hat{\Delta}$	bias	s.e.	MSE	$\hat{\Delta}_{R=1}$	bias	s.e.	MSE
IPW obs	1.000	0.000	0.087	0.008	0.998	-0.002	0.056	0.003
mean difference	0.882	-0.118	0.066	0.018	0.849	-0.151	0.064	0.027
true effect (normalized)	1.000				1.000			

Note: 1999 Monte Carlo replications. 'IPW obs' controls for attrition related to observables. All effects are normalized to 1.

Table 1 displays the estimates, bias, standard errors (s.e.) and mean squared errors (MSE) for the different methods when attrition is related to observables. Note that the ATEs on the entire population and on the respondents are normalized to unity. The IPW estimators following from Propositions 2 and 3 are effective in controlling for attrition bias. $\hat{\Delta}, \hat{\Delta}_{R=1}$ are close to the true values and their accuracy in terms of MSE increases in the sample size. In contrast, the mean difference estimator is substantially biased. For the estimation of the ATE on the entire population under the smaller sample size it is, however, competitive in terms of the MSE. This is due to its smaller variance compared to IPW, which introduces additional uncertainty through the estimation of two propensity scores. Under the larger sample size, the persistence of the bias

of the mean difference estimator dominates its better precision, such that IPW is clearly superior.

Many, and in particular a lot of recently conducted social experiments exceed the sample sizes considered in the simulations and typically contain several thousand observations, e.g., Angrist, Bettinger, and Kremer (2006), Angrist and Lavy (2009), Bertrand and Mullainathan (2004), Gertler (2004), and Krueger and Zhu (2004), or even more (Karlan and List, 2007). Simulation results suggest that IPW is likely to reduce the MSE under non-random attrition when using such sufficiently large data sets.

Table 2 displays the results for IPW estimators (i) controlling for attrition on unobservables ('IPW unobs', following from Propositions 4 and 5) and (ii) observables alone ('IPW obs', following from Propositions 2 and 3) when response depends also on unobservables. The former methods exploiting the instrument entail only moderate biases and MSEs, whereas the accuracy of IPW only controlling for attrition on observables is considerably lower. When estimating the ATE on the entire population, 'IPW obs' performs even worse than the mean difference estimator. This suggests that omitting important factors in a model for attrition may be worse than not controlling for response bias at all.

Table 2: Attrition on unobservables, simulated data

	n=500							
	$\hat{\Delta}$	bias	s.e.	MSE	$\hat{\Delta}_{R=1}$	bias	s.e.	MSE
IPW unobs	0.974	-0.026	0.118	0.015	1.009	0.009	0.110	0.012
IPW obs	0.780	-0.220	0.108	0.060	0.898	-0.102	0.098	0.020
mean difference	0.891	-0.109	0.118	0.026	0.878	-0.122	0.116	0.028
true effect (normalized)	1.000				1.000			
	n=2000							
	$\hat{\Delta}$	bias	s.e.	MSE	$\hat{\Delta}_{R=1}$	bias	s.e.	MSE
IPW unobs	0.979	-0.021	0.059	0.004	1.014	0.014	0.054	0.003
IPW obs	0.783	-0.217	0.055	0.050	0.900	-0.100	0.049	0.012
mean difference	0.892	-0.108	0.059	0.015	0.879	-0.121	0.058	0.018
true effect (normalized)	1.000				1.000			

Note: 1999 Monte Carlo replications. 'IPW unobs' controls for attrition related to observables and unobservables, 'IPW obs' only for attrition related to observables. All effects are normalized to 1.

To illustrate the potential gains of IPW in empirical applications when attrition is related to unobservables, we use a publicly available subsample of Tennessee's Project STAR Experiment for the subsequent simulation study based on empirical data. Project STAR was conducted in

the mid-1980s to evaluate the effects of small class sizes (target 13-17 students instead of 22-25 students in regular classes) in kindergartens and schools on student achievement, see for example Finn and Achilles (1990, 1999) and Krueger (1999). Our data set contains 5,852 children of which 1,757 were randomly assigned to a small class in kindergarten. The outcome of interest (Y) is the Stanford Achievement Test (SAT) in maths at the end of the kindergarten year (average test score: 485.377, standard deviation: 47.698).

A major issue for the applicability of the proposed IPW methods under attrition related to unobservables is the requirement of a continuous instrument. Therefore, future experimental designs might consider the inclusion of (close to) continuous instruments which should ideally be randomly assigned in a similar way as the treatment. As mentioned before, the number of phone calls could be used to instrument the response rate to post-treatment surveys. Up to date, however, such variables are typically not available in social experiments (see for instance Lee, 2009), and this is also the case for Project STAR. For this reason we will pursue a somewhat unorthodox simulation approach to investigate the performance of IPW when estimating the ATE on the entire population under attrition related to unobservables.

We discard all observations assigned to a small class in kindergarten and treat the sample of controls ($n = 4,095$) as if it was the entire population of interest. Moreover, we randomly assign a binary placebo treatment D among the controls with a 'treatment probability' of 0.5. Thus, the true treatment effect is known to be zero. Furthermore, the experiment is artificially broken by the introduction of the following response process, which is designed such that roughly two thirds of the outcomes are observed:

$$R_i \quad = \quad I\{\beta_0 + \beta_1 D_i + \beta_2 X_i - \beta_3 Z_i - \beta_4 U_i + V_i > 0\},$$

where $\beta_0 = 2.5, \beta_1 = \beta_2 = \beta_3 = 1, \beta_4 = 2$. X denotes race (one if white and zero otherwise) which we treat as being observed. U represents socio-economic status (one if eligible for free lunches, zero otherwise) and is assumed not to be observed for the sake of the subsequent simulation.

The only generated variables apart from the treatment are the instrument Z (uniform, support $[-1.5, 1.5]$) and the error term V (standard normal).

An inspection of the data shows that both X and U are strongly correlated with Y. Thus, neglecting attrition supposedly biases estimation. Note, however, that the exact structural relation between race, socio-economic status, and the SAT score is unknown due to the use of the empirical data. This is fundamentally different to the conventional Monte Carlo design (see the previous simulations) where the outcome equation is explicitly modeled. The motivation for using empirical data is to conduct simulations that are more closely linked to real world problems and, thus, more credible than studies merely based on generated data. Simulations relying on empirical data can also be found in Bertrand, Duflo, and Mullainathan (2004), Diamond and Sekhon (2006), Huber, Lechner, and Wunsch (2010b), and Lee and Whang (2010).

In each of the 1999 Monte Carlo replications, (Y, X, U) are randomly drawn from the 'population' without replacement and (D, Z, V) from their respective distributions in order to compute R. Then, we estimate the response propensity score by regressing R on $(1, D, X, Z)$ and the treatment propensity score, which is unknown in our simulation due to the use of empirical data, by regressing D on $(1, X, \hat{p}(W), X \times \hat{p}(W))$ using probit specifications. Contrary to the previous simulations, we estimate the effects with and without trimming of propensity scores as some values are close to the boundaries. We consider two trimming levels, where response propensity scores smaller than 5% (10%) and treatment propensity scores smaller than 5% (10%) or larger than 95 % (90%) are trimmed to the respective threshold values.

Table 3 presents the results which are in line with the previous simulations. The bias of the IPW estimator is moderate irrespective of the sample size and the trimming level, whereas it amounts to roughly 1.8 SAT scores when taking mean differences. Yet, when considering the smaller sample size, the mean difference estimator is superior w.r.t. the MSE as it is more precise than IPW. However, as the sample size increases IPW increasingly outperforms mean differences. We therefore conclude that weighting based on instruments is effective in reestablishing the validity of experiments under attrition on unobservables, at least when the sample size is not too

small such that the gain in bias reduction outweighs the loss in precision due to the estimation of the propensity scores and weighting. Of course, a precondition for this result is the availability of a continuous instrument that is both relevant (sufficiently correlated with response behavior) and valid (no direct effect on the outcome). Table 3 also reports the average numbers of trimmed response and treatment propensity scores in the simulations, which are moderate even for the 10% trimming level.

Table 3: Attrition on unobservables, simulated data, zero treatment effect

	n=500				
	$\hat{\Delta}$	bias	s.e.	MSE	av. num. of trimmed p, π
IPW unobs (untrimmed)	0.425	0.425	6.197	38.585	
IPW unobs (5% trimming)	0.428	0.428	6.198	38.594	0.000, 16.191
IPW unobs (10% trimming)	0.443	0.443	6.200	38.631	0.452, 46.177
mean difference	1.768	1.768	5.420	32.389	
true effect	0.000				
n=2000					
	$\hat{\Delta}$	bias	s.e.	MSE	av. num. of trimmed p, π
IPW unobs (untrimmed)	0.411	0.411	2.994	9.134	
IPW unobs (5% trimming)	0.412	0.412	2.994	9.137	0.000, 24.315
IPW unobs (10% trimming)	0.423	0.423	2.997	9.158	0.012, 142.857
mean difference	1.772	1.772	2.667	10.252	
true effect	0.000				

Note: 1999 Monte Carlo replications.

'IPW unobs' controls for attrition related to observables and unobservables.

6 Empirical application

We present an application of Propositions 2 and 3 (attrition related to observables) to a large data set coming from a randomized experiment conducted to assess the Mexican universal health insurance program 'Seguro Popular'. These data have been collected, analyzed, and provided by King, Gakidou, Imai, Lakin, Moore, Nall, Ravishankar, Vargas, Tellez-Rojo, Hernandez-Avila, Hernandez-Avila, and Hernandez-Llamas (2009), see also King, Gakidou, Ravishankar, Moore, Lakin, Vargas, Tellez-Rojo, Hernandez-Avila, Hernandez-Avila, and Hernandez-Llamas (2007) for more details concerning the experimental design. According to the authors, one major goal of Mexico's health insurance reform in 2003 was to provide quality care and health coverage also to poor households, in particular to the (at that time) 50 million uninsured Mexicans. The

target after completion of the roll-out was to increase total health spending in Mexico by a full percentage point of the GDP (from 5.6% in 2002).

The data set provides information on 32,515 households, including socio-economic characteristics (gender of the interviewee, education, household size, wealth, employment, etc.), health insurance coverage, detailed health information (both physical and mental), use of and spending on health services, and many other factors. One member of each household was interviewed shortly before the random assignment took place and roughly 10 months after the baseline survey to measure the outcomes of interest. Note that the program assignment was randomized on the regional level. 100 geographical health clusters were matched to each other in order to form similar pairs w.r.t. important observed covariates. Then, randomization took place within cluster pairs resulting in the matched-pair, cluster-randomization design discussed in Imai, King, and Nall (2009).

In the subsequent discussion, we focus on the program's impact on symptoms reflecting mental health problems. A priori, we would expect the program to increase mental health for two reasons. Firstly, we suspect that health coverage increases the recipients' probability to claim and receive treatment against whatsoever health deficiencies. Secondly, the program may cover health expenditures previously paid out of the households' own pockets. By easing the financial burden the program may also reduce stress, anxiety, and other mental symptoms related to the necessity to spend a substantial share of household income or wealth on health. The follow-up survey contains four questions related to mental health that are measured on a scale from 1 (none) to 5 (extreme): sleeping problems, not feeling rested and fresh during the day, sadness or depression, worry or anxiety.

As shown in Table 4, taking mean differences (and controlling for the fact that randomization is clustered when computing standard errors) suggests that the program has some positive impact on mental health, as it significantly reduces the severeness of the first, third, and fourth symptom. However, the outcomes are only available for 28,015 individuals (86 % of those in the baseline survey), as 4,500 did not answer the questions of interest in the follow-up survey. We therefore

estimate the ATEs on respondents and on the entire population under the assumption of attrition related to observables to check the robustness of the estimates. To be precise, we actually focus on the intention to treat effect (ITT). I.e., we do not consider and/or control for non-compliance w.r.t. the program assignment, which is beyond the scope of this paper.

Table 4: Means and mean differences in mental health outcomes

In the last 30 days, how much of a problem did you have	av. treated	av. non-tr.	diff.	p-value
...with sleeping	1.486	1.552	-0.067	0.036
...due to not feeling rested during the day	1.558	1.637	-0.079	0.038
...with feeling sad, low or depressed	1.740	1.808	-0.068	0.548
...with worry or anxiety	1.479	1.588	-0.108	0.051

Note: Standard errors account for clustering.

Estimating the ATE on respondents ($\Delta_{R=1}$) requires us to control for all variables that affect the mental health outcomes and are not balanced across the treatment states conditional on response. It is well acknowledged in the literature that socio-economic variables such as education, age, occupation, household composition, wealth, and income are strongly correlated with health (see for instance Llena-Nozal, Lindeboom, and Portrait, 2004, and Mulatu and Schooler, 2002). Furthermore, the pre-treatment health state (see Huber, Lechner, and Wunsch 2010) as well as health behavior (e.g., alcohol consumption, smoking, possession of a health insurance plan, visits to receive health care) are most likely further important confounders. Even regional factors such as living in a rural or urban area and ethnicity might influence mental health.

Running a probit regression of the treatment state on the aforementioned characteristics reveals that only the possession of a health insurance plan prior to the treatment differs significantly (at the 5% level) across treated and nontreated respondents. The other variables are well balanced. Table 5 presents the estimates for $\Delta_{R=1}$ when conditioning on health insurance in the treatment propensity score. The estimator is based on the normalized sample analog of Proposition 2 without trimming, as the problem of extreme propensity score values does not occur. The effects are only slightly smaller than the mean differences presented in Table 4 and by no means significantly different from the latter. The standard errors are computed based on block bootstrapping with 1999 replications which accounts for cluster randomization.

To estimate the ATE on the entire population (Δ), we check whether the aforementioned variables that are likely to affect the health outcomes are also related to the response behavior. This is indeed the case for gender, age, education, ethnicity, household size, health behavior, and living in a rural/urban area. We therefore condition on these variables in a very flexible way (by including dummies and interaction terms) when estimating the response propensity score. The probit specification is provided in Appendix A.8. Interestingly, the coefficients do not differ significantly across treatment states. This indicates that the differences in the observed characteristics of respondents and the entire population are comparable conditional on $D = 1$ and $D = 0$, see the discussion of Assumption 3.

We use the normalized sample analog of Proposition 3, again without trimming, to estimate the ATE. For the first, second, and fourth outcome (those with significant mean differences), the estimates are somewhat smaller (effects decrease between 6.4% and 7.5% in absolute terms) than the effects in Table 4. While being significant on the 5% level, the estimates are, however, not significantly different from the mean differences. Thus, the ATE of the health insurance program 'Seguro Popular' on mental health does not appear to vary importantly across the different covariate distributions of the respondents and the entire population.

Table 5: ATEs on respondents and the entire population

In the last 30 days, how much of a problem did you have	$\hat{\Delta}_{R=1}$	p-value	$\hat{\Delta}$	p-value
...with sleeping	-0.065	0.041	-0.062	0.031
...due to not feeling rested during the day	-0.076	0.045	-0.074	0.026
...with feeling sad, low or depressed	-0.071	0.530	-0.057	0.508
...with worry or anxiety	-0.106	0.052	-0.101	0.006

Note: p-values are computed based on 1999 block bootstraps.
Attrition is assumed to depend on observables.

7 Conclusion

This paper discusses the identification of treatment effects in randomized experiments when outcomes are only partially observed due to attrition and non-response in follow-up surveys. Its first contribution is the systematic coverage of various forms of attrition, i.e., when outcomes are

missing completely at random and when attrition is related to observables (missing at random) and unobservables. We treat these various forms by imposing different assumptions on the relation between the response behavior and the treatment, the observed covariates, and the unobserved characteristics in a fairly general treatment effect model.

The second contribution is to show point identification of average treatment effects on the respondents and on the entire population based on different implementations of inverse probability weighting (IPW). Each IPW method is tailored to a specific nature of attrition considered, which provides practitioners with straightforward solutions depending on the suspected missing data problem. In particular, we introduce an IPW approach based on an instrumental variable (IV) strategy to tackle attrition on unobservables, which was not considered in the experimental literature before.

Despite its technical ease of implementation, the IV-based IPW approach appears to be rarely applicable in social experiments conducted up to date, due to the lack of credible continuous instruments for attrition. This is unfortunate, as attrition on unobservables seems to be a potential threat in many fields of research where randomized trials take place (such as education and labor economics), in particular when the number of observed baseline characteristics is low. Future research might therefore consider the creation and random assignment of instruments in experimental designs in order to validate and increase the credibility of experimental inference.

A Appendix

A.1 Proof of Proposition 1

Under Assumptions 3 and 3', the ATE is identified by

$$\Delta \quad = \quad E\left[\frac{R\cdot Y}{p(1,X)}|D=1\right]-E\left[\frac{R\cdot Y}{p(0,X)}|D=0\right].$$

Proof:

$$\begin{aligned}
& E\left[\frac{R\cdot Y}{p(1,X)}|D=1\right]-E\left[\frac{R\cdot Y}{p(0,X)}|D=0\right] \\
= \quad & E\left[E\left[\frac{R\cdot Y}{p(1,X)}|X=x,D=1\right]|D=1\right]-E\left[E\left[\frac{R\cdot Y}{p(0,X)}|X=x,D=0\right]|D=0\right] \\
= \quad & E\left[E\left[\frac{Y}{p(1,X)}|R=1,X=x,D=1\right]\cdot p(1,X)|D=1\right] \\
- \quad & E\left[E\left[\frac{Y}{p(0,X)}|R=1,X=x,D=0\right]\cdot p(0,X)|D=0\right] \\
= \quad & E\left[E\left[Y|R=1,X=x,D=1\right]|D=1\right]-E\left[E\left[Y|R=1,X=x,D=0\right]|D=0\right] \\
= \quad & E\left[E\left[Y|X=x,D=1\right]|D=1\right]-E\left[E\left[Y|X=x,D=0\right]|D=0\right] \\
= \quad & E\left[Y|D=1\right]-E\left[Y|D=0\right]=E\left[Y^1\right]-E\left[Y^0\right]=\Delta.
\end{aligned}$$

The first equality follows from the law of iterated expectations, the fourth from Assumption 3, implying that Y and R are independent conditional on (X,D). The fifth follows from a backward application of the law of iterated expectations, and the sixth equality follows from the randomization of the treatment.

A.2 Proof of Proposition 2

Under Assumptions 4 and 4', the ATE on respondents is identified by

$$\Delta_{R=1} \quad = \quad E\left[\frac{D\cdot Y}{\pi(X)}-\frac{(1-D)\cdot Y}{1-\pi(X)}|R=1\right].$$

Proof:

$$
\begin{aligned}
& E\left[\frac{D\cdot Y}{\pi(X)}-\frac{(1-D)\cdot Y}{1-\pi(X)}|R=1\right] \\
= \; & E\left[E\left[\frac{D\cdot Y}{\pi(X)}|X=x,R=1\right]|R=1\right]-E\left[E\left[\frac{(1-D)\cdot Y}{1-\pi(X)}|X=x,R=1\right]|R=1\right] \\
= \; & E\left[E\left[\frac{Y}{\pi(X)}|D=1,X=x,R=1\right]\cdot\pi(X)|R=1\right] \\
- \; & E\left[E\left[\frac{Y}{1-\pi(X)}|D=0,X=x,R=1\right]\cdot(1-\pi(X))|R=1\right] \\
= \; & E\left[E\left[Y|D=1,X=x,R=1\right]|R=1\right]-E\left[E\left[Y|D=0,X=x,R=1\right]|R=1\right] \\
= \; & E\left[E\left[Y|D=1,X=x\right]|R=1\right]-E\left[E\left[Y|D=0,X=x\right]|R=1\right] \\
= \; & E\left[E\left[Y^1|X=x\right]|R=1\right]-E\left[E\left[Y^0|X=x\right]|R=1\right] \\
= \; & E\left[Y^1|R=1\right]-E\left[Y^0|R=1\right]=\Delta_{R=1}.
\end{aligned}
$$

The first equality follows from the law of iterated expectations, the fourth from Assumption 4, implying that Y and R are independent conditional on (X,D). The fifth follows from the randomization of the treatment and the sixth equality follows from a backward application of the law of iterated expectations.

A.3 Proof of Proposition 3

Under Assumptions 3', 4, and 4', the ATE is identified by

$$
\Delta \;=\; E\left[\frac{R\cdot D\cdot Y}{p(D,X)\cdot\pi(X)}-\frac{R\cdot(1-D)\cdot Y}{p(D,X)\cdot(1-\pi(X))}\right].
$$

Proof:

$$
\begin{aligned}
& E\left[\frac{R\cdot D\cdot Y}{p(D,X)\cdot\pi(X)}-\frac{R\cdot(1-D)\cdot Y}{p(D,X)\cdot(1-\pi(X))}\right] \\
= \; & E\left[E\left[\frac{R\cdot D\cdot Y}{p(D,X)\cdot\pi(X)}|X=x\right]\right]-E\left[E\left[\frac{R\cdot(1-D)\cdot Y}{p(D,X)\cdot(1-\pi(X))}|X=x\right]\right] \\
= \; & E\left[E\left[\frac{Y}{p(D,X)\cdot\pi(X)}|R=1,D=1,X=x\right]\cdot p(D,X)\cdot\pi(X)\right] \\
- \; & E\left[E\left[\frac{Y}{p(D,X)\cdot(1-\pi(X))}|R=1,D=0,X=x\right]\cdot p(D,X)\cdot(1-\pi(X))\right] \\
= \; & E\left[E\left[Y|R=1,D=1,X=x\right]\right]-E\left[E\left[Y|R=1,D=0,X=x\right]\right] \\
= \; & E\left[E\left[Y|D=1,X=x\right]\right]-E\left[E\left[Y|D=0,X=x\right]\right] \\
= \; & E\left[E\left[Y^1|X=x\right]\right]-E\left[E\left[Y^0|X=x\right]\right] \\
= \; & E\left[Y^1\right]-E\left[Y^0\right]=\Delta.
\end{aligned}
$$

The first equality follows from the law of iterated expectations, the fourth from Assumption 4, implying that Y and R are independent conditional on (X, D). The fifth follows from the randomization of the treatment and the sixth equality follows from a backward application of the law of iterated expectations.

A.4 Proof of Proposition 4

Under Assumptions 5, 5', and (5"b), the ATE on the respondents is identified by

$$\Delta_{R=1} \quad = \quad E\left[\frac{D \cdot Y}{\pi(X, p(W))} \mid R=1\right] - E\left[\frac{(1-D) \cdot Y}{1-\pi(X, p(W))} \mid R=1\right].$$

Proof:

$$\begin{aligned}
& E\left[\frac{D \cdot Y}{\pi(X, p(W))} \mid R=1\right] - E\left[\frac{(1-D) \cdot Y}{(1-\pi(X, p(W)))} \mid R=1\right] \\
= & \underset{p(W)}{E}\left[\underset{X}{E}\left[E\left[\frac{D \cdot Y}{\pi(X, p(W))} - \frac{(1-D) \cdot Y}{(1-\pi(X, p(W)))} | X, p(W), R=1\right] | p(W), R=1\right] \mid R=1\right] \\
= & \underset{p(W)}{E}\left[\underset{X}{E}\left[E\left[\frac{Y}{\pi(X, p(W))} | D=1, X, p(W), R=1\right] \cdot \pi(X, p(W))\right.\right. \\
- & \left.\left. E\left[\frac{Y}{(1-\pi(X, p(W)))} | D=0, X, p(W), R=1\right] \cdot (1-\pi(X, p(W))) | p(W), R=1\right] \mid R=1\right] \\
= & \underset{p(W)}{E}\left[\underset{X}{E}\left[E\left[Y | D=1, X, p(W), R=1\right] - E\left[Y | D=0, X, p(W), R=1\right] | p(W), R=1\right] \mid R=1\right] \\
= & \underset{p(W)}{E}\left[\underset{X}{E}\left[E\left[Y^1 | X, p(W), R=1\right] - E\left[Y^0 | X, p(W), R=1\right] | p(W), R=1\right] \mid R=1\right] \\
= & \underset{p(W)}{E}\left[\underset{X}{E}\left[\Delta_{R=1}(X, p(W)) | p(W), R=1\right] \mid R=1\right] = \Delta_{R=1}.
\end{aligned}$$

The first equality follows from the law of iterated expectations, the fourth from Assumptions 5 and 5'. $\Delta_{R=1}(X, p(W))$ denotes the conditional ATE given X and $p(W)$ in the selected subpopulation. Finally, the last equality is a backward application of the law of iterated expectations.

A.5 Proof of Proposition 5

Under Assumptions 5, 5', and 5", the ATE is identified by

$$\Delta \quad = \quad E\left[\frac{R \cdot D \cdot Y}{p(W) \cdot \pi(X, p(W))}\right] - E\left[\frac{R \cdot (1-D) \cdot Y}{p(W) \cdot (1-\pi(X, p(W)))}\right]. \tag{1}$$

Proof:

$$
\begin{aligned}
& E\left[\frac{R\cdot D\cdot Y}{p(W)\cdot\pi(X,p(W))}\right]-E\left[\frac{R\cdot(1-D)\cdot Y}{p(W)\cdot(1-\pi(X,p(W)))}\right] \\
= & \underset{p(W)}{E}\left[\underset{X}{E}\left[E\left[\frac{R\cdot D\cdot Y}{p(W)\cdot\pi(X,p(W))}-\frac{R\cdot(1-D)\cdot Y}{p(W)\cdot(1-\pi(X,p(W)))}|X,p(W)\right]|p(W)\right]\right] \\
= & \underset{p(W)}{E}\left[\underset{X}{E}\left[E\left[\frac{D\cdot Y}{p(W)\cdot\pi(X,p(W))}-\frac{(1-D)\cdot Y}{p(W)\cdot(1-\pi(X,p(W)))}|R=1,X,p(W)\right]\cdot p(W)|p(W)\right]\right] \\
= & \underset{p(W)}{E}\left[\underset{X}{E}\left[E\left[\frac{D\cdot Y}{\pi(X,p(W))}-\frac{(1-D)\cdot Y}{(1-\pi(X,p(W)))}|R=1,X,p(W)\right]|p(W)\right]\right] \\
= & \underset{p(W)}{E}\left[\underset{X}{E}\left[E\left[\frac{Y}{\pi(X,p(W))}|D=1,R=1,X,p(W)\right]\cdot\pi(X,p(W))\right.\right. \\
- & \left.\left.E\left[\frac{Y}{(1-\pi(X,p(W)))}|D=0,R=1,X,p(W)\right]\cdot(1-\pi(X,p(W)))|p(W)\right]\right] \\
= & \underset{p(W)}{E}\left[\underset{X}{E}\left[E\left[Y|D=1,R=1,X,p(W)\right]-E\left[Y|D=0,R=1,X,p(W)\right]|p(W)\right]\right] \\
= & \underset{p(W)}{E}\left[\underset{X}{E}\left[E\left[Y^1|R=1,X,p(W)\right]-E\left[Y^0|R=1,X,p(W)\right]|p(W)\right]\right] \\
= & \underset{p(W)}{E}\left[\underset{X}{E}\left[\Delta_{R=1}(X,p(W))|p(W)\right]\right]=\underset{p(W)}{E}\left[\underset{X}{E}\left[\Delta(X,p(W))|p(W)\right]\right]=\Delta,
\end{aligned}
$$

The first equality follows from the law of iterated expectations, the sixth from Assumptions 5 and 5'. The eighth equality follows from Assumption (5'c) by which $F_{U|D=d,X=x,p(W)=p(w),R=r}=F_{U|X=x,p(W)=p(w),R=r}$ and Assumption (5"c) which imposes additivity of observed and unobserved terms. Both together imply that $\Delta_{R=1}(X,p(W))$, the conditional ATE given X and $p(W)$ among respondents, is equal to $\Delta_{R=0}(X,p(W))$ and thus, $\Delta(X,p(W))$. Finally, the last equality is a backward application of the law of iterated expectations.

A.6 Proof of Proposition 6

Under Assumptions 6, 6', and (6"b), the ATE on the respondents is identified by

$$
\Delta_{R=1} \quad = \quad E\left[\frac{D\cdot Y}{\pi(p(D,Z))}\mid R=1\right]-E\left[\frac{(1-D)\cdot Y}{1-\pi(p(D,Z))}\mid R=1\right].
$$

Proof:

$$
\begin{aligned}
& E\left[\frac{D\cdot Y}{\pi(p(D,Z))} \mid R=1\right] - E\left[\frac{(1-D)\cdot Y}{(1-\pi(p(D,Z)))} \mid R=1\right] \\
= \;& E\left[E\left[\frac{D\cdot Y}{\pi(p(D,Z))} - \frac{(1-D)\cdot Y}{(1-\pi(p(D,Z)))}|p(D,Z), R=1\right] \mid R=1\right] \\
= \;& E\Big[E\left[\frac{Y}{\pi(p(D,Z))}|D=1,p(D,Z),R=1\right]\cdot\pi(p(D,Z)) \\
- \;& E\left[\frac{Y}{(1-\pi(p(D,Z)))}|D=0,p(D,Z),R=1\right]\cdot(1-\pi(p(D,Z))) \mid R=1\Big] \\
= \;& E\left[E\left[Y|D=1,p(D,Z),R=1\right] - E\left[Y|D=0,p(D,Z),R=1\right] \mid R=1\right] \\
= \;& E\left[E\left[Y^1|p(D,Z),R=1\right] - E\left[Y^0|p(D,Z),R=1\right] \mid R=1\right] \\
= \;& E\left[\Delta_{R=1}(p(D,Z)) \mid R=1\right] = \Delta_{R=1}.
\end{aligned}
$$

The first equality follows from the law of iterated expectations, the fourth from Assumptions 6 and 6'. $\Delta_{R=1}(X, p(W))$ denotes the conditional ATE given $p(D,Z)$ in the selected subpopulation. Finally, the last equality is a backward application of the law of iterated expectations.

A.7 Proof of Proposition 7

Under Assumptions 6, 6', and 6", the ATE is identified by

$$
\Delta = E\left[\frac{R\cdot D\cdot Y}{p(D,Z)\cdot\pi(p(D,Z))}\right] - E\left[\frac{R\cdot(1-D)\cdot Y}{p(D,Z)\cdot(1-\pi(p(D,Z)))}\right]. \tag{2}
$$

Proof:

$$
\begin{aligned}
& E\left[\frac{R\cdot D\cdot Y}{p(D,Z)\cdot\pi(p(D,Z))}\right] - E\left[\frac{R\cdot(1-D)\cdot Y}{p(D,Z)\cdot(1-\pi(p(D,Z)))}\right] \\
= \;& E\left[E\left[\frac{R\cdot D\cdot Y}{p(D,Z)\cdot\pi(p(D,Z))} - \frac{R\cdot(1-D)\cdot Y}{p(D,Z)\cdot(1-\pi(p(D,Z)))}|p(D,Z)\right]\right] \\
= \;& E\left[E\left[\frac{D\cdot Y}{p(D,Z)\cdot\pi(p(D,Z))} - \frac{(1-D)\cdot Y}{p(D,Z)\cdot(1-\pi(p(D,Z)))}|R=1,p(D,Z)\right]\cdot p(D,Z)\right] \\
= \;& E\left[E\left[\frac{D\cdot Y}{\pi(p(D,Z))} - \frac{(1-D)\cdot Y}{(1-\pi(p(D,Z)))}|R=1,p(D,Z)\right]\right] \\
= \;& E\Big[E\left[\frac{Y}{\pi(p(D,Z))}|D=1,R=1,p(D,Z)\right]\cdot\pi(p(D,Z)) \\
- \;& E\left[\frac{Y}{(1-\pi(p(D,Z)))}|D=0,R=1,p(D,Z)\right]\cdot(1-\pi(p(D,Z)))\Big] \\
= \;& E\Big[E\left[Y|D=1,R=1,p(D,Z)\right] - E\left[Y|D=0,R=1,p(D,Z)\right]\Big] \\
= \;& E\left[E\left[Y^1|R=1,p(D,Z)\right] - E\left[Y^0|R=1,p(D,Z)\right]\right] \\
= \;& E\left[\Delta_{R=1}(p(D,Z))|p(D,Z)\right] = E\left[\Delta(p(D,Z))|p(D,Z)\right] = \Delta,
\end{aligned}
$$

The first equality follows from the law of iterated expectations, the sixth from Assumptions 6 and 6'. The eighth equality follows from Assumption (6'c) by which $\Delta_{R=1}(p(D,Z))$, the conditional ATE given $p(D,Z)$ among respondents, is equal to $\Delta_{R=0}(X,p(D,Z))$ and thus, $\Delta(p(D,Z))$. Finally, the last equality is a backward application of the law of iterated expectations.

A.8 Specification of the response propensity score

Table 6: Probit specification of the response propensity score

variable	coefficient	(s.e.)	p-value
age	0.069	(0.012)	0.000
age squared	-0.001	(0.000)	0.000
age < 20	0.812	(0.257)	0.002
age btw. 20 and 25	0.489	(0.201)	0.015
age btw. 26 and 30	0.539	(0.162)	0.001
age btw. 31 and 35	0.446	(0.129)	0.001
age btw. 36 and 40	0.346	(0.100)	0.001
age btw. 41 and 45	0.198	(0.085)	0.020
age btw. 46 and 50	0.126	(0.075)	0.091
age btw. 51 and 55	0.095	(0.072)	0.187
missing age	1.794	(0.462)	0.000
gender	0.069	(0.076)	0.368
missing gender	-0.764	(0.430)	0.076
indigenous	0.107	(0.062)	0.082
missing ethnicity	-0.390	(0.156)	0.013
education: secondary	-0.074	(0.026)	0.005
education: preparatory/vocational	-0.106	(0.037)	0.004
education: normal	-0.251	(0.062)	0.000
education: technical/commercial	-0.255	(0.083)	0.002
missing education	-0.224	(0.160)	0.162
household size: 1	-0.369	(0.055)	0.000
household size: 2	-0.189	(0.032)	0.000
household size: 3	-0.101	(0.025)	0.000
household educ: secondary	-0.098	(0.040)	0.015
household educ: prep./voc.	-0.300	(0.065)	0.000
household educ: techn./comm.	-0.217	(0.167)	0.193
missing household educ	-0.033	(0.055)	0.550
ever outpatient health care	0.122	(0.028)	0.000
smoker	-0.083	(0.035)	0.019
missing smoker	-0.069	(0.088)	0.435
using seatbelt when in a car	-0.215	(0.050)	0.000
missing seatbelt	-0.069	(0.036)	0.057
urban/rural	0.552	(0.108)	0.000
gender*age < 20	-0.224	(0.093)	0.015
gender*age btw. 20 and 25	-0.187	(0.080)	0.019
gender*age btw. 26 and 30	-0.467	(0.082)	0.000
gender*age btw. 31 and 35	-0.424	(0.082)	0.000
gender*age btw. 36 and 40	-0.346	(0.077)	0.000
gender*age btw. 41 and 45	-0.299	(0.082)	0.000
gender*age btw. 46 and 50	-0.214	(0.079)	0.007
gender*age btw. 51 and 55	-0.211	(0.088)	0.016
gender*using seatbelt	0.147	(0.050)	0.003
gender*urban/rural	-0.176	(0.066)	0.008
using seatbelt*age < 20	0.240	(0.127)	0.058
urban/rural*age < 20	-0.376	(0.098)	0.000
constant	-1.077	(0.427)	0.012

Note: Standard errors (s.e.) account for clustering. Pseudo-R^2=0.081.

References

Ahn, H., and J. Powell (1993): "Semiparametric Estimation of Censored Selection Models with a Nonparametric Selection Mechanism," *Journal of Econometrics*, 58, 3–29.

Angrist, J., E. Bettinger, and M. Kremer (2006): "Long-Term Educational Consequences of Secondary School Vouchers: Evidence from Administrative Records in Colombia," *American Economic Review*, 96, 847–862.

Angrist, J., and V. Lavy (2009): "The Effects of High Stakes High School Achievement Awards: Evidence from a Randomized Trial," *The American Economic Review*, 99, 1384–1414.

Barnard, J., C. E. Frangakis, J. L. Hill, and D. B. Rubin (2003): "Principal Stratification Approach to Broken Randomized Experiments: A Case Study of School Choice Vouchers in New York City," *Journal of the American Statistical Association*, 98, 299–311.

Bertrand, M., E. Duflo, and S. Mullainathan (2004): "How Much Should We Trust Differences-in-Differences Estimates?," *The Quarterly Journal of Economics*, 119, 249–275.

Bertrand, M., and S. Mullainathan (2004): "Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination," *The American Economic Review*, 94, 991–1013.

Bloom, H., L. Orr, S. Bell, G. Cave, F. Doolittle, W. Lin, and J. Bos (1997): "The Benefits and Costs of JTPA Title II-A Programs: Key Findings from the National Job Training Partnership Act Study," *Journal of Human Resources*, 32, 549–576.

Blundell, R., and J. Powell (2003): "Endogeneity in Nonparametric and Semiparametric Regression Models," in *Advances in Economics and Econometrics*, ed. by L. H. M. Dewatripont, and S. Turnovsky, pp. 312–357. Cambridge University Press, Cambridge.

Busso, M., J. DiNardo, and J. McCrary (2009a): "Finite Sample Properties of Semiparametric Estimators of Average Treatment Effects," *unpublished manuscript*.

——— (2009b): "New Evidence on the Finite Sample Properties of Propensity Score Matching and Reweighting Estimators," *IZA Discussion Paper No. 3998*.

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

Cochran, W. G., and S. P. Chambers (1965): "The planning of observational studies of human populations," *Journal of the Royal Statistical Society Series A*, 128, 234–265.

Crump, R. K., V. J. Hotz, G. W. Imbens, and O. A. Mitnik (2009): "Dealing with limited overlap in estimation of average treatment effects," *Biometrika*, 96, 187–199.

Das, M., W. K. Newey, and F. Vella (2003): "Nonparametric Estimation of Sample Selection Models," *Review of Economic Studies*, 70, 33–58.

Diamond, A., and J. S. Sekhon (2006): "Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies," *Institute of Governmental Studies Working Paper*.

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

Duflo, E. (2006): "Field Experiments in Development Economics," *unpublished manuscript*.

Finn, J. D., and C. M. Achilles (1990): "Answers and Questions about Class Size: A Statewide Experiment," *American Educational Research Journal*, 27, 557–577.

——— (1999): "Tennessee's Class Size Study: Findings, Implications, Misconceptions," *Educational Evaluation and Policy Analysis*, 21, 97–109.

Fisher, R. (1925): *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh.

——— (1935): *The Design of Experiments*. Oliver and Boyd, Edinburgh.

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

Frangakis, C. E., and D. B. Rubin (2002): "The defining role of principal stratification and effects for comparing treatments adjusted for posttreatment variables: from treatment noncompliance to surrogate endpoints," *Biometrics*, 58, 191199.

Freedman, D. (2006): "Statistical Models for Causation: What Inferential Leverage Do They Provide," *Evaluation Review*, 30, 691–713.

Gertler, P. (2004): "Do Conditional Cash Transfers Improve Child Health? Evidence from PROGRESA's Control Randomized Experiment," *The American Economic Review*, 94, 336–341.

Grilo, C. M., R. Money, D. H. Barlow, A. W. Goddard, J. M. Gorman, S. G. Hofmann, L. A. Papp, M. K. Shear, and S. W. Woods (1998): "Pretreatment patient factors predicting attrition from a multicenter randomized controlled treatment study for panic disorder," *Comprehensive Psychiatry*, 39, 323–332.

Grogger, J. (2009): "Bounding the Effects of Social Experiments: Accounting for Attrition in Administrative Data," *mimeo*.

Harrison, G. W., and J. A. List (2004): "Field Experiments," *Journal of Economic Literature*, 42(1009-1055).

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

Heckman, J. J. (1974): "Shadow Prices, Market Wages and Labor Supply," *Econometrica*, 42, 679–694.

——— (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.

——— (1979): "Sample selection bias as a specification error," *Econometrica*, 47, 153–161.

Heitjan, D. F., and S. Basu (1996): "Distinguishing Missing at Random and Missing Completely at Random," *The American Statisticia*, 50, 207–213.

Hirano, K., G. W. Imbens, and G. Ridder (2003): "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica*, 71, 1161–1189.

Horowitz, J., and C. F. Manski (1998): "Censoring of outcomes and regressors due to survey nonresponse: Identification and estimation using weights and imputations," *Journal of Econometrics*, 84, 37–58.

——— (2000): "Nonparametric Analysis of Randomized Experiments with Missing Covariate and Outcome Data," *Journal of the American Statistical Association*, 95, 77–84.

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

Huber, M. (2009): "Treatment evaluation in the presence of sample selection," *Discussion Paper, Department of Economics, University of St. Gallen*, 09-07.

Huber, M., M. Lechner, and C. Wunsch (2010a): "Does leaving welfare improve health? Evidence for Germany," *forthcoming in Health Economics*.

——— (2010b): "Reliable estimators for propensity score matching estimation," *mimeo*.

Imai, K. (2009): "Statistical analysis of randomized experiments with non-ignorable missing binary outcomes: an application to a voting experiment," *Journal of the Royal Statistical Society Series C*, 58, 83–104.

Imai, K., G. King, and C. Nall (2009): "The Essential Role of Pair Matching in Cluster-Randomized Experiments, with Application to the Mexican Universal Health Insurance Evaluation," *Statistical Science*, (24), 29–53.

Imbens, G. W. (2004): "Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review," *The Review of Economics and Statistics*, 86, 4–29.

——— (2009): "Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009)," *NBER Working Paper No. 14896*.

Imbens, G. W., and W. Newey (2003): "Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity," *Econometrica*, 77, 1481–1512.

Imbens, G. W., and J. M. Wooldridge (2009): "Recent Developments in the Econometrics of Program Evaluation," *Journal of Economic Literature*, 47, 5–86.

Karlan, D., and J. A. List (2007): "Does Price Matter in Charitable Giving? Evidence from a Large-Scale Natural Field Experiment," *The American Economic Review*, 97, 1774–1793.

King, G., E. Gakidou, K. Imai, J. Lakin, R. T. Moore, C. Nall, N. Ravishankar, M. Vargas, M. M. Tellez-Rojo, J. E. Hernandez-Avila, M. Hernandez-Avila, and H. Hernandez-Llamas (2009): "Public policy for the poor? A randomised assessment of the Mexican universal health insurance programme," *Lancet*, 373, 1447–1454.

King, G., E. Gakidou, N. Ravishankar, R. T. Moore, J. Lakin, M. Vargas, M. M. Tellez-Rojo, J. E. Hernandez-Avila, M. Hernandez-Avila, and H. Hernandez-Llamas (2007): "A 'politically robust' experimental design for public policy evaluation, with application to the Mexican Universal Health Insurance program," *Journal of Policy Analysis and Management*, 26, 479–506.

Krueger, A. B. (1999): "Experimental Estimates of Education Production Functions," *Quarterly Journal of Economics*, 114, 497–532.

Krueger, A. B., and P. Zhu (2004): "Another Look at the New York City School Voucher Experiment," *American Behavioral Scientist*, 47, 658–698.

Lee, D. S. (2009): "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects," *Review of Economic Studies*, 76, 1071–1102.

Lee, S., and Y.-J. Whang (2010): "Nonparametric tests of conditional treatment effects," *cemmap Working Paper CWP36/09*.

Llena-Nozal, A., M. Lindeboom, and F. Portrait (1045-1062): "The effect of work on mental health: does occupation matter?," *Health Economics*, 13.

Mealli, F., and B. Pacini (2008): "Causal inference with nonignorably missing outcomes: instrumental variables and principal stratification," *mimeo*.

Mulatu, S., and C. Schooler (2002): "Causal Connections between Socio-Economic Status and Health: Reciprocal Effects and Mediating Mechanisms," *Journal of Health and Social Behavior*, 43, 22–41.

Newey, W., J. Powell, and F. Vella (1999): "Nonparametric Estimation of Triangular Simultaneous Equations Models," *Econometrica*, 67, 565–603.

Newey, W. K. (2007): "Nonparametric continuous/discrete choice models," *International Economic Review*, 48, 1429–1439.

Neyman, J. (1923): "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles.," *Statistical Science*, Reprint, 5, 463–480.

Robins, J., and A. Rotnitzky (1995): "Semiparametric Efficiency in Multivariate Regression Models with Missing Data," *Journal of American Statistical Association*, 90, 122–129.

Robins, J., A. Rotnitzky, and L. Zhao (1995): "Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data," *Journal of American Statistical Association*, 90, 106–121.

Robins, J., and A. Tsiatis (1991): "Correcting for Non-Compliance in Randomized Trials Using Rank-Preserving Structural Failure Time Models," *Communications in Statistics*, 20, 2069–2631.

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

Rotnitzky, A., and J. Robins (1995): "Semiparametric regression estimation in the presence of dependent censoring," *Biometrika*, 82, 805–820.

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

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

——— (1977): "Formalizing Subjective Notions About the Effect of Nonrespondents in Sample Surveys," *Journal of the American Statistical Association*, 72, 538–543.

——— (1978): "Multiple Imputations in Sample Surveys-A Phenomenological Bayesian Approach to Nonresponse," in *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 20–34.

——— (1990): "Formal Modes of Statistical Inference For Causal Effects," *Journal of Statistical Planning and Inference*, 25, 279–292.

——— (1996): "Multiple Imputation After 18+ Years," *Journal of the American Statistical Association*, 91, 473–489.

——— (2008): "For objective causal inference, design trumps analysis," *The Annals of Applied Statistics*, 2, 808–840.

Scharfstein, D. O., A. Rotnitzky, and J. M. Robins (1999): "Adjusting for Nonignorable Drop-Out Using Semiparametric Nonresponse Models," *Journal of the American Statistical Association*, 94, 1096–1120.

Wooldridge, J. (2002): "Inverse Probability Weigthed M-Estimators for Sample Selection, Attrition and Stratification," *Portuguese Economic Journal*, 1, 141–162.

——— (2007): "Inverse probability weighted estimation for general missing data problems," *Journal of Econometrics*, 141, 1281–1301.

Zhang, J., and D. B. Rubin (2003): "Estimation of causal effects via principal stratification when some outcome are truncated by death," *Journal of Educational and Behavioral Statistics*, 28, 353–368.

Zhang, J., D. B. Rubin, and F. Mealli (2008): "Evaluating The Effects of Job Training Programs on Wages through Principal Stratification," in *Advances in Econometrics: Modelling and Evaluating Treatment Effects in Econometrics*, ed. by D. Millimet, J. Smith, and E. Vytlacil, vol. 21, pp. 117–145. Elsevier Science Ltd.