OXFORD BULLETIN OF ECONOMICS AND STATISTICS, 77, 1 (2015) 0305–9049
doi: 10.1111/obes.12056

Sharp Bounds on Causal Effects under Sample Selection*

Martin Huber and Giovanni Mellace

Department of Economics, University of St. Gallen, Varnbuelstrasse 14, CH 9000, St. Gallen Switzerland (e-mail: martin.huber@unisg.ch; giovanni.mellace@unisg.ch)

Abstract

In many empirical problems, the evaluation of treatment effects is complicated by sample selection so that the outcome is only observed for a non-random subpopulation. In the absence of instruments and/or tight parametric assumptions, treatment effects are not point identified, but can be bounded under mild restrictions. Previous work on partial identification has primarily focused on the 'always observed' (irrespective of the treatment). This article complements those studies by considering further populations, namely the 'compliers' (observed only if treated) and the observed population. We derive sharp bounds under various assumptions and provide an empirical application to a school voucher experiment.

I. Introduction

The sample selection problem, see for instance Gronau (1974) and Heckman (1974), arises when the outcome of interest is only observed for a non-randomly selected subpopulation. This may flaw causal analysis and is an ubiquitous phenomenon in many fields where treatment effect evaluations are conducted, such as labour, health and educational economics. For example, in the estimation of the returns to a training, it is an issue when only a selective subgroup of training participants and non-participants finds employment which is a condition for observing earnings. Similar problems are inherent in clinical trials when some of the participants in medical treatments pass away ('truncation by death') before the health outcome is measured. As a further example, consider the effect of randomly provided private schooling on college entrance examinations. The sample selection problem arises if only a non-random subgroup of students takes the exam.

In sample selection models in economics (see the seminal work of Heckman, 1974, 1976, 1979), identification commonly relies on tight functional form restrictions and the availability of a valid instrument for selection. Albeit the literature has recently moved towards more flexible models, see for instance Das *et al.* (2003) and Newey (2009), it

*We have benefited from comments by Michael Lechner, Fabrizia Mealli, Franco Peracchi, Christoph Rothe, seminar/conference participants in St. Gallen (research seminar) and Pisa (4th Italian Congress of Econometrics and Empirical Economics), and two anonymous referees.

JEL Classification numbers: C14, C21, C24.

typically imposes strong assumptions on the unobserved terms unlikely to hold in many applications, see Huber and Melly (2012), or uses invalid instruments, see Huber and Mellace (2013). Similar arguments apply to many studies in the related field of missing data problems, which often use regression or weighting adjustments (assuming selection on observables) to control for missing outcomes, see for instance Hausman and Wise (1979), Robins *et al.* (1995) and Wooldridge (2007). In the absence of unattractive parametric restrictions or instruments for sample selection, treatment effects are not point identified, but upper and lower bounds can still be obtained under fairly mild restrictions.

Partial identification of economic parameters in general goes back to Manski (1989, 1994) and Robins (1989). In the context of the sample selection (or missing outcome data) problem, several contributions in the fields of principal stratification, see Frangakis and Rubin (2002) and econometrics derive bounds on treatment effects. Building on Horowitz and Manski (1998), Horowitz and Manski (2000) consider the partial identification of the average treatment effect in the entire population assuming a binary outcome and also allowing for missing covariate information. Zhang and Rubin (2003) (see also Zhang, Rubin and Mealli, 2008) bound the average treatment effects for one subpopulation, namely the 'always observed', whose outcomes are observed both under treatment and non-treatment. They impose two assumptions both separately and jointly: (i) monotonicity of selection in the treatment and (ii) stochastic dominance of the potential outcomes of the always observed over those of other populations. Imai (2008) shows that the bounds of Zhang and Rubin (2003) are sharp and additionally considers the identification of quantile treatment effects. Lee (2009) invokes monotonicity of selection (but not stochastic dominance) when assessing the average earnings effects of Job Corps, a training program for disadvantaged youths in the US, on the always observed and proves the sharpness of the bounds. Blanco *et al.* (2011) evaluate the same program, but add assumptions on the order of mean potential outcomes within and across subpopulations to obtain tighter bounds. In contrast to the aforementioned contributions, Lechner and Melly (2007) bound the effects on those treated *and* observed, which is a mixed population consisting of always observed and 'compliers' who are observed under treatment, but would not be without treatment.[1]

The main contribution of this article is the derivation of sharp bounds (the tightest feasible bounds given the assumptions imposed and the information available) on average treatment effects among compliers, 'defiers' (outcomes observed if not treated and not observed if treated) and the observed population, which have not been considered in previous work. We show that under the monotonicity and/or stochastic dominance assumptions, informative bounds can be derived even when the outcomes of particular subpopulations are only observed in one treatment arm. For instance, one useful result is that under both assumptions, the lower bound on the observed population coincides with that on the always observed. This is relevant for many applications where particular interest lies in whether the lower bound includes a zero effect. Thus, the assumptions may bear considerable identifying power, which is demonstrated in an application to a school voucher experiment in Colombia previously analyzed by Angrist *et al.* (2006).

[1] This definition is not to be confused with the local average treatment effect framework (see Imbens and Angrist (1994)), where compliers are those who are treated if assigned to treatment and not treated if assigned to control in a randomized trial.

We argue that it is, depending on the evaluation problem at hand, useful to look at further target populations than the ones covered in the literature so far, which can be done by the methods proposed in this article. For example, it is the compliers whose selection state reacts on the treatment and it appears interesting in many applications whether this is observed along with (and may be rooted in) a particular treatment effect. Taking the wage effects of a job training program as an example, one might want to learn whether the change in the employment state due to the training is accompanied by an increase in the potential wage. If yes, this points to an increase in productivity that may be at least partly responsible for finding employment. Furthermore, in particular applications, the compliers might also bear more policy relevance than the always observed. For example, consider a school voucher experiment investigating the effect of private schooling on test scores in a college entrance exam which are only observed conditional on taking the exam. As the compliers do so only under private schooling, they are likely to come from educationally more disadvantaged families than the always observed. This might exactly be the group policy makers want to target. As a further example which is related to truncation by death, assume that a medical treatment reduces mortality but has detrimental side effects on health. In this particular set-up, the always observed are clearly not of policy interest: as these individuals always survive, one would in any case not expose them to a treatment that may harm their health state. In contrast, the compliers are those who survive thanks to the treatment and are therefore highly relevant. For this reason, one may want to assess the magnitude of the adverse effect on this group, for example, for developing alternative treatments that are less harmful.

Furthermore, we might prefer to make causal statements rather for larger shares of the entire population than for smaller groups. The largest possible group for which at least one potential outcome (under treatment or non-treatment) is observed constitutes the observed population, which is again a mixture of several subpopulations. For example, policy makers might want to learn about the average effects on all individuals whose outcomes are observed, without thinking in terms of different latent subpopulations. Also, Newey (2007) considers this population, however, investigating point identification based on continuous instruments. As an example, one might prefer to evaluate the returns to training for those working, who often make up a substantial share of the entire population. Note that in some cases, the observed population may even be more relevant than the entire population, because the latter also includes the never observed for which the evaluation of causal effects does not always appear useful. For example, under truncation by death, the potential health states of individuals that never survive (under any treatment) are not defined and therefore, this group appears irrelevant when evaluating a medical treatment.[2] In contrast, one might want to learn about the health effects on all survivors. Likewise, one could challenge the relevance of assessing the effects of a schooling intervention (like school vouchers) on college entrance examinations in the entire population, given that the latter also includes individuals who would never take these exams or enter college. In contrast, we may be interested in the effects on those participating in the exams and thus expressing the desire to go to college.

[2] Note, however, that mortality could be a relevant outcome on its own.

Finally, there is also a statistical argument to look at populations other than the always observed. In fact, if neither monotonicity nor unattractive parametric assumptions are imposed and if the share of 'never observed' (whose outcomes are not observed irrespective of the treatment state) is larger than the one of the always observed, no informative bounds can be obtained for the latter. However, informative (albeit generally quite large) bounds are still available for the observed population.

The remainder of this article is organized as follows. Section II formally characterizes the sample selection problem based on principal stratification. Section III discusses partial identification of treatment effects for the compliers and the observed population under no assumptions (worst case bounds) as well as under monotonicity and/or stochastic dominance. Estimators are presented in section IV. An empirical application to a school voucher experiment in Colombia is provided in section V. Section VI concludes.

II. The selection problem

Assume that we are interested in the effect of a binary treatment, $T_i \in \{1, 0\}$, on an outcome Y_i some time after assignment. Using the potential outcome framework, see for instance Neyman (1923), Fisher (1935), and Rubin (1977), we will denote by $Y_i(1)$ and $Y_i(0)$ the two potential outcomes that individual i would receive under treatment and non-treatment. Even under randomization of the treatment, post-treatment complications might introduce selection bias and flaw causal inference. One particular form of such complications is sample selection, implying that the outcome of interest is only observed for a non-random subpopulation. To address this problem, let $S_i \in \{1, 0\}$ be an observed binary post-treatment selection indicator which is 1 if the outcome of some individual is observed and 0 otherwise. Furthermore, we denote by $S_i(1)$ and $S_i(0)$ the two potential selection states. Then, we can express the selection indicator and the observed outcome as functions of the respective potential states:

$$S_i = T_i \cdot S_i(1) + (1 - T_i) \cdot S_i(0),$$
$$Y_i = T_i \cdot Y_i(1) + (1 - T_i) \cdot Y_i(0) \quad \text{if } S_i = 1 \text{ and not observed otherwise.}$$

That is, at best (if $S_i = 1$) one of the two potential outcomes is observed. As at least one potential outcome remains unknown, both point and partial identification of treatment effects require further assumptions. The first restriction maintained throughout the discussion is the so-called Stable Unit Treatment Value Assumption (SUTVA, e.g. Rubin, 1990), which rules out interference between units and general equilibrium effects of the treatment:

Assumption 1.

$$Y_i(t) \perp T_j \, \forall j \neq i,$$
$$S_i(t) \perp T_j \, \forall j \neq i.$$

'$\perp$' denotes independence. This implies that the potential post-treatment variables of any subject i are unrelated to the treatment status of any other individual.

Causal inference requires the specification of the treatment assignment mechanism. If randomly assigned, the treatment is independent of the potential values of the post-treatment variables S, Y. However, in many observational studies, randomization is assumed to hold

TABLE 1

Principal strata

Principal strata (G_i)	$S_i(1)$	$S_i(0)$	*Appellation*
11	1	1	Always observed
10	1	0	Compliers
01	0	1	Defiers
00	0	0	Never observed

only conditional on some observed pre-treatment covariates X. This assumption is known as conditional independence assumption (CIA), also referred to as 'selection on observables' or 'unconfoundedness', see for instance Imbens (2004) and Imbens and Wooldridge (2009). As elsewhere in the sample selection literature, see for instance Lee (2009) and Mealli and Pacini (2008), we therefore assume that the joint distribution of the potential outcomes and selection states is independent of the treatment given X:[3]

Assumption 2.

$$T_i \perp (S(1)_i, S_i(0), Y_i(1), Y_i(0)) \,|\, X_i = x, \quad \forall x \in \mathcal{X},$$

where $\mathcal{X}$ denotes the support of X. In the further discussion, conditioning on X will be kept implicit. Therefore, all results either refer to the experimental framework, see also the application further below, or to an analysis within cells with the same values of X.

As shown in Table 1 and discussed in Zhang and Rubin (2003), the population can be divided into four subpopulations or principal strata (denoted as G), according to the value the selection indicator $S_i(t)$ takes under different treatment states.

The terms 'always/never observed', 'compliers' and 'defiers' are in the spirit of Imbens and Angrist (1994) and Angrist *et al.* (1996), who, however, consider the conceptually different problem of treatment endogeneity. Assumption 2 implies that the stratum G some individual belongs to is independent of the treatment assignment and that the potential outcomes are independent of the treatment conditional on the principal stratum. Therefore, any treatment effect defined within a principal stratum is a well defined causal parameter. The problem for identification is that either $S_i(1)$ or $S_i(0)$ but never both are known for any individual so that the principal stratum to which a subject belongs is not directly observed. Without further assumptions, neither the principal strata proportions nor the distributions of the potential outcomes within each stratum are identified. To see this, note that the observed values of T_i and S_i generate the following four observed subgroups, denoted as $o(T_i, S_i)$, which are all mixtures of two principal strata (Table 2).

Therefore, also the probability to belong to an observed subgroup is a mixture of principal strata proportions, henceforth denoted as $\pi_{ss'} \equiv \Pr(S(1) = s, S(0) = s')$. Let $P_{s|t}$ represent the observed selection probability conditional on treatment, $\Pr(S = s \,|\, T = t)$, in the population of interest. Under Assumption 2, which ensures that the strata proportions conditional on the treatment are equal to the unconditional strata proportions, the relation between the observed $P_{s|t}$ and the latent $\pi_{ss'}$ is as follows.

[3] Assumption 2 may be replaced by $T_i \perp (S_i(1), S_i(0), Y_i(0)) | X_i = x \; \forall x \in \mathcal{X}$, if the inference is conditional on $T = 1$, that is, if one is only interested in treatment effects on the treated.

TABLE 2

Observed subgroups and principal strata

Observed subgroups $o(T_i, S_i)$	*Principal strata*	Y_i *observed*
$o(1,1) = \{i : T_i = 1, S_i = 1\}$	Subject i belongs either to 11 or to 10	Yes
$o(1,0) = \{i : T_i = 1, S_i = 0\}$	Subject i belongs either to 01 or to 00	No
$o(0,1) = \{i : T_i = 0, S_i = 1\}$	Subject i belongs either to 11 or to 01	Yes
$o(0,0) = \{i : T_i = 0, S_i = 0\}$	Subject i belongs either to 10 or to 00	No

Thus, point identification of causal effects can only be obtained by imposing unattractive parametric assumptions, see for instance the discussion in Mealli and Pacini (2008), Zhang *et al.* (2009), and Heckman (1974, 1976, 1979). However, intervals of treatment effects for particular strata that are consistent with the observed data can be derived under milder assumptions. As mentioned before, treated and non-treated units are only observed for the always observed (stratum 11), that is, those observed irrespective of the treatment state. For this reason, most of the literature on bounding treatment effects under sample selection focuses on stratum 11, see Zhang and Rubin (2003), Grilli and Mealli (2008), Zhang *et al.* (2008) and Lee (2009), with the exception of Lechner and Melly (2007).

We, however, argue that the always observed are generally not the only population of interest and show that informative bounds can also be derived for other populations under assumptions which seem plausible in many applications. In particular, we are interested in the effects in stratum 10 and in the entire observed population ($S = 1$). Stratum 10 consists of those individuals observed with and not observed without treatment. Thus, they can be referred to as 'compliers' in selection with respect to the treatment. This stratum is interesting in many applications as it consists of the marginal population that changes the selection state due to the treatment. Taking the wage effects of a job training program as an example, we might be interested in whether the change in the employment state due to the training is accompanied by an increase in the potential outcomes. If yes, this points to an increase in productivity that may be at least partly responsible for finding employment. Furthermore, in a school voucher experiment investing the effect of private schooling on college entrance examinations, the compliers are those who take the test only under private schooling. They are therefore likely to be more disadvantaged and academically less challenged at home than the always observed, which may be exactly the population policy makers want to target.

The observed population is a mixture of always observed, compliers and defiers (stratum 01: observed under non-treatment, not observed under treatment) and therefore encounters individuals with different 'selection behaviours'. Still, policy makers might want to learn about the effects on all individuals whose outcomes are observed irrespective of their stratum affiliations. After all, the observed population contains all subjects for which at least one outcome (under treatment and/or non-treatment) is observed so that reasonable bounds on the effects may still be attained under the assumptions discussed below. In contrast, for the never observed bounds are most likely very uninformative and in addition, this group does not always appear policy relevant (e.g. under truncation by death, this is the group that never survives so that their outcomes are not defined). Thus, the observed group appears to be the largest possible subpopulation for which useful inference appears to be feasible.

To link principal stratification to the econometric literature on sample selection, we conclude this section by discussing the identification problem in a structural model (see also Mealli and Pacini, 2008; Mellace and Rocci, 2011; Huber, 2012):

$$\begin{aligned} Y_i &= \varphi(T_i, U_i), \\ S_i &= I\{\varsigma(T_i, V_i) \geqslant 0\}, \\ T_i &= I\{\psi(\zeta_i) \geqslant 0\}, \end{aligned} \tag{1}$$

where $I\{\cdot\}$ is the indicator function, φ, ς, ψ are unknown functions and U, V, ζ are unobserved terms. $\zeta \perp U, V$ by random assignment (or conditional on X by the conditional independence assumption in observational studies). The selectivity of S depends on the relationship of the unobserved terms U and V. Note that the sample selection problem disappears when conditioning on V because then, S and U are conditionally independent. Even though V is unknown, the problem can be controlled for if there exists a function $G(V)$ so that

$$U \perp S \mid G(V).$$

Imbens (2006) calls such a function 'type of unit'. Principal stratification is a natural choice of $G(\cdot)$, as

$$\begin{aligned} G(v) = G(v') \quad &\text{if} \quad \varsigma(t, v) = \varsigma(t, v') \quad \forall t, v \neq v', \\ G(v) \neq G(v') \quad &\text{if} \quad \varsigma(t, v) \neq \varsigma(t, v') \quad \text{for some } t, v \neq v', \end{aligned}$$

and $U \perp S | G(V)$ by construction. Once we condition on the 'type of unit', selection becomes ignorable. Principal stratification represents the coarsest possible choice of the type function. As pointed out by Imbens (2006), the optimal type function is any functional that is constant on sets of values of V which, for all values t, lead to the same value of S.

III. Assumptions and partial identification

Worst case bounds

We restrict the support of the potential outcomes to be bounded: $Y_i(1), Y_i(0) \in \mathcal{Y} \equiv [Y^{\text{LB}}, Y^{\text{UB}}]$, where $-\infty < Y^{\text{LB}} < Y^{\text{UB}} < \infty$ are the values at the lower and upper end of the support $\mathcal{Y}$, respectively. This condition rules out infinite upper or lower bounds on the average treatment effects (ATE) in any population even without imposing restrictions other than Assumptions 1 and 2. We will also assume that Y is continuous, whereas the adaptation of our methods to discrete outcomes is discussed in the Supporting Information.

Partial identification is obtained in three steps. In the first step, we derive sharp bounds on the principal strata proportions using Assumption 2. As one can express three out of four proportions as a function of the remaining one, we only need to bound the latter. Therefore, all bounds are computed as functions of the defier proportion, but choosing any other principal stratum would entail the same results. The second step (which is mostly discussed in the Supporting Information) gives the bounds on the mean potential outcomes and the ATEs conditional on the defier proportion. It makes use of the fact that each observed conditional mean outcome is a mixture of the potential outcome distributions of two principal strata, with the mixing probabilities corresponding to the relative principal strata proportions:

TABLE 3

Observed probabilities and principal strata proportions

Observed conditional selection probabilities	*Principal strata proportions*
$P_{1\|1} \equiv \Pr(S=1\|T=1)$	$\pi_{11}+\pi_{10}$
$P_{0\|1} \equiv \Pr(S=0\|T=1)$	$\pi_{01}+\pi_{00}$
$P_{1\|0} \equiv \Pr(S=1\|T=0)$	$\pi_{11}+\pi_{01}$
$P_{0\|0} \equiv \Pr(S=0\|T=0)$	$\pi_{10}+\pi_{00}$

$$f(Y|T=0,S=1)=\frac{\pi_{11}}{\pi_{11}+\pi_{01}}\cdot f(Y(0)|G=11)+\frac{\pi_{01}}{\pi_{11}+\pi_{01}}\cdot f(Y(0)|G=01) \quad (2)$$

and

$$f(Y|T=1,S=1)=\frac{\pi_{11}}{\pi_{11}+\pi_{10}}\cdot f(Y(1)|G=11)+\frac{\pi_{10}}{\pi_{11}+\pi_{10}}\cdot f(Y(1)|G=10), \quad (3)$$

Given the defier proportion (and thus, the mixing probabilities, in our case $\pi_{11}/(\pi_{11}+\pi_{01})$, $\pi_{01}/(\pi_{11}+\pi_{01})$ and $\pi_{11}/(\pi_{11}+\pi_{10})$, $\pi_{10}/(\pi_{11}+\pi_{10})$, respectively), the results of Horowitz and Manski (1995) (see section 3.2 and proposition 4 therein) can be directly applied to derive sharp bounds on the mean potential outcomes, as well as any functional of the potential outcomes that respects stochastic dominance (e.g. quantiles), in equations (2) and (3). Finally, maximizing the ATEs in the second step over admissible defier proportions that satisfy Assumption 2 yields the sharp upper (lower) bounds on the ATEs. Similarly, one can bound other parameters such as quantile treatment effects.[4]

Concerning the bounds on the defier proportion, note that under Assumptions 1 and 2, Table 3 provides us with the following equations:

$$\begin{aligned} P_{1|0}-\pi_{01}=\pi_{11} &\Rightarrow \pi_{01}\leqslant P_{1|0},\\ P_{0|1}-\pi_{01}=\pi_{00} &\Rightarrow \pi_{01}\leqslant P_{0|1},\\ P_{1|1}-P_{1|0}+\pi_{01}=\pi_{10} &\Rightarrow \pi_{01}\geqslant P_{1|0}-P_{1|1}, \end{aligned}$$

so that

$$\pi_{01}\in[\max(0,P_{1|0}-P_{1|1}),\ \min(P_{1|0},P_{0|1})]. \quad (4)$$

In the absence of further assumptions, Zhang and Rubin (2003) derive the worst case bounds of the ATE on the always observed (stratum 11), henceforth denoted as Δ_{11}, which are shown to be sharp in Imai (2008). For the sake of brevity, let $\bar{Y}_{t,s}\equiv E(Y|T=t,S=s)$, that is, the mean of Y given $T=t$ and $S=s$ (which is only observed for $S=1$). Furthermore, let $F_{Y_{t,s}}(y)\equiv \Pr(Y\leqslant y|T=t,S=s)$ and $F^{-1}_{Y_{t,s}}(q)\equiv \inf\{y: F_{Y_{t,s}}(y)\geqslant q\}$, that is, the conditional cumulated distribution function and quantile function of Y given $T=t$ and $S=s$. Finally, let $\bar{Y}_{t,s}(\min|q)\equiv E(Y|T=t,S=s,Y\leqslant F^{-1}_{Y_{t,s}}(q))$ and $\bar{Y}_{t,s}(\max|q)\equiv E(Y|T=t,S=s,Y\geqslant F^{-1}_{Y_{t,s}}(1-q))$. The upper and the lower bound of the ATE on the always observed $\Delta_{11}\equiv E(Y(1)-Y(0)|G=11)$, denoted as $\Delta^{\mathrm{UB}}_{11}$ and $\Delta^{\mathrm{LB}}_{11}$, in Zhang and Rubin (2003), are

[4] One can adapt the results of Stoye (2010) to our framework to also bound spread parameters, for example, variances of the treatment effects.

$$\begin{aligned}\Delta_{11}^{\mathrm{UB}} &= \min_{\pi_{01}}[\bar{Y}_{1,1}(\max|(P_{1|0}-\pi_{01})/P_{1|1}) - \bar{Y}_{0,1}(\min|(P_{1|0}-\pi_{01})/P_{1|0})],\\ \Delta_{11}^{\mathrm{LB}} &= \max_{\pi_{01}}[\bar{Y}_{1,1}(\min|(P_{1|0}-\pi_{01})/P_{1|1}) - \bar{Y}_{0,1}(\max|(P_{1|0}-\pi_{01})/P_{1|0})].\end{aligned} \tag{5}$$

Thus, the authors suggest to optimize over all possible values of the defiers' share π_{01} that are consistent with the data to obtain the upper and lower bound. A first contribution of the present work is to show that numerical optimization is not necessary. As outlined in the the Supporting Information, $\Delta_{11}^{\mathrm{UB}}$ and $\Delta_{11}^{\mathrm{LB}}$ can be simplified to

$$\begin{aligned}\Delta_{11}^{\mathrm{UB}} &= \bar{Y}_{1,1}(\max|(P_{1|0}-\pi_{01}^{\max})/P_{1|1}) - \bar{Y}_{0,1}(\min|(P_{1|0}-\pi_{01}^{\max})/P_{1|0}),\\ \Delta_{11}^{\mathrm{LB}} &= \bar{Y}_{1,1}(\min|(P_{1|0}-\pi_{01}^{\max})/P_{1|1}) - \bar{Y}_{0,1}(\max|(P_{1|0}-\pi_{01}^{\max})/P_{1|0}),\end{aligned} \tag{6}$$

where $\pi_{01}^{\max} \equiv \min(P_{1|0}, P_{0|1})$, as the highest feasible defiers' share maximizes the upper bound and minimizes the lower bound. Note that the bounds are only informative (i.e. tighter than $Y^{\mathrm{UB}} - Y^{\mathrm{LB}}$) if $P_{1|0} > P_{0|1}$, which has also been noticed by Lee (2009). This implies that $\pi_{11} > \pi_{00}$, that is, that the share of always observed is larger than the share of never observed. In this case, $(P_{1|0} - \pi_{01}^{\max})/P_{1|1} = (P_{1|0} - P_{0|1})/P_{1|1}$ and $(P_{1|0} - \pi_{01}^{\max})/P_{1|0} = (P_{1|0} - P_{0|1})/P_{1|0}$, so that the bounds only depend on this ratio of observed proportions.

In contrast to previous work, we will now also derive bounds for the compliers (stratum 10), the defiers (stratum 01), and the observed population. It is obvious from our previous discussion that the share of compliers in $o(1,1)$ is $\pi_{10}/(\pi_{11} + \pi_{10}) = (P_{1|1} - P_{1|0} + \pi_{01})/P_{1|1}$, that is, the fraction of those who are not always observed. This allows us to bound the upper and lower values of the mean potential outcome under treatment by $\bar{Y}_{1,1}(\max|(P_{1|1} - P_{1|0} + \pi_{01}^{\min})/P_{1|1})$ and $\bar{Y}_{1,1}(\min|(P_{1|1} - P_{1|0} + \pi_{01}^{\min})/P_{1|1})$, respectively, where $\pi_{01}^{\min} \equiv \max(0, P_{1|0} - P_{1|1})$. However, nothing can be said about the mean potential outcome under non-treatment, as there are no compliers in $o(0,1)$. This requires us to assume the theoretical upper and lower bounds of the outcome Y^{UB} and Y^{LB}. Then, the sharp upper and lower bounds of the ATE on the compliers $\Delta_{10} \equiv E(Y(1) - Y(0)|G = 10)$, denoted as $\Delta_{10}^{\mathrm{UB}}$ and $\Delta_{10}^{\mathrm{LB}}$, are

$$\begin{aligned}\Delta_{10}^{\mathrm{UB}} &= \bar{Y}_{1,1}(\max|(P_{1|1} - P_{1|0} + \pi_{01}^{\min})/P_{1|1}) - Y^{\mathrm{LB}},\\ \Delta_{10}^{\mathrm{LB}} &= \bar{Y}_{1,1}(\min|(P_{1|1} - P_{1|0} + \pi_{01}^{\min})/P_{1|1}) - Y^{\mathrm{UB}}.\end{aligned} \tag{7}$$

These bounds are informative only if $P_{1|0} - P_{1|1} < 0 \Rightarrow \pi_{10} > \pi_{01}$ (more compliers than defiers). Then, $(P_{1|1} - P_{1|0} + \pi_{01}^{\min})/P_{1|1} = (P_{1|1} - P_{1|0})/P_{1|1}$. The proofs for the sharpness of these and all other bounds proposed below are provided in the Supporting Information.

Similarly, the share of defiers in $o(0,1)$ is $\pi_{01}/(\pi_{11} + \pi_{01}) = \pi_{01}/P_{1|0}$. This allows us to bound the upper and lower value of the mean potential outcome under non-treatment by $\bar{Y}_{0,1}(\max|\pi_{01}^{\min}/P_{1|0})$ and $\bar{Y}_{0,1}(\min|\pi_{01}^{\min}/P_{1|0})$, respectively. As there are no defiers in $o(1,1)$, we again need to invoke Y^{UB} and Y^{LB}. The sharp upper and lower bounds for the ATE on the defiers $\Delta_{01} \equiv E(Y(1) - Y(0)|G = 01)$, denoted as $\Delta_{01}^{\mathrm{UB}}$ and $\Delta_{01}^{\mathrm{LB}}$, are

$$\begin{aligned}\Delta_{01}^{\mathrm{UB}} &= Y^{\mathrm{UB}} - \bar{Y}_{0,1}(\min|\pi_{01}^{\min}/P_{1|0}),\\ \Delta_{01}^{\mathrm{LB}} &= Y^{\mathrm{LB}} - \bar{Y}_{0,1}(\max|\pi_{01}^{\min}/P_{1|0}).\end{aligned} \tag{8}$$

These bounds are only informative if $P_{1|0} - P_{1|1} > 0 \Rightarrow \pi_{01} > \pi_{10}$, that is, if the defiers' share is at least as large as the compliers' share. If this is true, then $\pi_{01}^{\min}/P_{1|1} = (P_{1|0} - P_{1|1})/P_{1|1}$.

This is, together with the identification result for the compliers, an interesting finding because it implies that without imposing monotonicity of selection in the treatment (as outlined below), bounds are informative either for the defiers or for the compliers, but never for both populations. It also implies that unless $P_{1|1} - P_{1|0} = 0$, either positive (if $P_{1|1} - P_{1|0} > 0$) or negative (if $P_{1|0} - P_{1|1} > 0$) monotonicity of S in T is consistent with the data, but not both at the same time. See the discussion in the next subsubsection.

Finally, we derive the worst case bounds for the ATE on the observed population $\Delta_{S=1} \equiv E(Y(1) - Y(0)|S = 1)$, which is a mixed population of always observed, compliers, and defiers. As

$$\begin{aligned}\Pr(S=1) &= \Pr(S=1|T=1)\cdot\Pr(T=1)+\Pr(S=1|T=0)\cdot\Pr(T=0)\\ &= \pi_{11}+\Pr(T=1)\cdot\pi_{10}+\Pr(T=0)\cdot\pi_{01},\end{aligned}$$

their respective shares in the observed population are given by

$$\begin{aligned}\frac{\pi_{11}}{\pi_{11}+\Pr(T=1)\cdot\pi_{10}+\Pr(T=0)\cdot\pi_{01}} &= \frac{(P_{1|0}-\pi_{01})}{\Pr(S=1)},\\ \frac{\Pr(T=1)\cdot\pi_{10}}{\pi_{11}+\Pr(T=1)\cdot\pi_{10}+\Pr(T=0)\cdot\pi_{01}} &= \frac{\Pr(T=1)\cdot(P_{1|1}-P_{1|0}+\pi_{01})}{\Pr(S=1)},\\ \frac{\Pr(T=0)\cdot\pi_{01}}{\pi_{11}+\Pr(T=1)\cdot\pi_{10}+\Pr(T=0)\cdot\pi_{01}} &= \frac{\Pr(T=0)\cdot\pi_{01}}{\Pr(S=1)}.\end{aligned}$$

In the Supporting Information, we use this fact to show that the bounds on $\Delta_{S=1}$ are given by

$$\begin{aligned}\Delta^{\mathrm{UB}}_{S=1} = &\frac{(P_{1|0}-\pi^{\max}_{01})\cdot\Pr(T=0)\cdot\bar{Y}_{1,1}(\max|b(P_{1|0}-\pi^{\max}_{01})/P_{1|1})}{\Pr(S=1)}\\ &-\frac{(P_{1|0}-\pi^{\max}_{01})\cdot\Pr(T=1)\cdot\bar{Y}_{0,1}(\min|(P_{1|0}-\pi^{\max}_{01})/P_{1|0})}{\Pr(S=1)}\\ &+\frac{\Pr(T=1)\cdot P_{1|1}\cdot\bar{Y}_{1,1}-\Pr(T=0)\cdot P_{1|0}\cdot\bar{Y}_{0,1}}{\Pr(S=1)}\\ &+\frac{\Pr(T=0)\cdot\pi^{\max}_{01}\cdot Y^{\mathrm{UB}}-\Pr(T=1)\cdot(P_{1|1}-P_{1|0}+\pi^{\max}_{01})\cdot Y^{\mathrm{LB}}}{\Pr(S=1)},\end{aligned} \tag{9}$$

$$\begin{aligned}\Delta^{\mathrm{LB}}_{S=1} = &\frac{(P_{1|0}-\pi^{\max}_{01})\cdot\Pr(T=0)\cdot\bar{Y}_{1,1}(\min|(P_{1|0}-\pi^{\max}_{01})/P_{1|1})}{\Pr(S=1)}\\ &-\frac{(P_{1|0}-\pi^{\max}_{01})\cdot\Pr(T=1)\cdot\bar{Y}_{0,1}(\max|(P_{1|0}-\pi^{\max}_{01})/P_{1|0})}{\Pr(S=1)}\\ &+\frac{\Pr(T=1)\cdot P_{1|1}\cdot\bar{Y}_{1,1}-\Pr(T=0)\cdot P_{1|0}\cdot\bar{Y}_{0,1}}{\Pr(S=1)}\\ &+\frac{\Pr(T=0)\cdot\pi^{\max}_{01}\cdot Y^{\mathrm{LB}}-\Pr(T=1)\cdot(P_{1|1}-P_{1|0}+\pi^{\max}_{01})\cdot Y^{\mathrm{UB}}}{\Pr(S=1)}.\end{aligned} \tag{10}$$

The identification region shrinks as the shares of compliers and/or defiers decreases. In the special case that both shares are zero the ATE on the observed population is point

identified. If the share of only one population is equal to zero, the bounds are equivalent to those under monotonicity which we will derive in the next subsubsection. Another result worth noting is that the bounds on the observed population are always informative. For example, if $P_{1|0} < P_{0|1}$, the bounds become

$$\Delta^{\mathrm{UB}}_{S=1} = \frac{\Pr(T=1)\cdot P_{1|1}}{\Pr(S=1)}\cdot(\bar{Y}_{1,1} - Y^{\mathrm{LB}}) + \frac{\Pr(T=0)\cdot P_{1|0}}{\Pr(S=1)}\cdot(Y^{\mathrm{UB}} - \bar{Y}_{0,1})$$

and

$$\Delta^{\mathrm{LB}}_{S=1} = \frac{\Pr(T=1)\cdot P_{1|1}}{\Pr(S=1)}\cdot(\bar{Y}_{1,1} - Y^{\mathrm{UB}}) + \frac{\Pr(T=0)\cdot P_{1|0}}{\Pr(S=1)}\cdot(Y^{\mathrm{LB}} - \bar{Y}_{0,1}),$$

which happen to be equal to the bounds under $\pi_{11} = 0$, that is, in the absence of always observed. Thus, even though informative bounds cannot be derived for the always observed if $P_{1|0} < P_{0|1}$, they can still be derived for the observed population.

Monotonicity

A commonly imposed assumption in the literature on partial identification of treatment effects under sample selection is weak monotonicity of selection with respect to the treatment:

Assumption 3. $\Pr(S_i(1) \geqslant S_i(0)) = 1$(monotonicity of selection).
In terms of the structural model in equation (1) this can be stated as

Assumption 3SM. $\varsigma(1, V_i) \geqslant \varsigma(0, V_i)\ \forall$ subjects i.

The monotonicity assumption requires that the potential selection state never decreases in the treatment and, thus, rules out the existence of the defiers (stratum 01). A symmetric result is obtained by assuming $\Pr(S_i(0) \geqslant S_i(1)) = 1$ which implies that stratum 10 does not exist. As already mentioned before, assuming $\Pr(S_i(1) \geqslant S_i(0)) = 1$ (positive monotonicity) is only consistent with the data if $P_{1|1} - P_{1|0} \geqslant 0$ and $\Pr(S_i(0) \geqslant S_i(1)) = 1$ (negative monotonicity) if $P_{1|0} - P_{1|1} \geqslant 0$. These are necessary, albeit not sufficient conditions for the respective monotonicity assumption. For the sake of brevity and due to the symmetry of the argumentation, we will only focus on Assumption 3 (positive monotonicity) in the subsequent discussion.

The plausibility of monotonicity depends on the empirical context. For example, it is not necessarily satisfied in the evaluation of the returns to a job training. In fact, employment (S) might react negatively on the training (T) due to reduced job search effort while being trained, a phenomenon known as 'lock-in' effect. Monotonicity might therefore only be plausible in later periods after the accomplishment of the training. The assumption seems more innocuous when evaluating the effectiveness of private schooling on college entrance examinations, given that private schooling offers a better education than public alternatives and affects the preferences for academic achievement. It appears reasonable to assume that students are more likely to take the test when receiving better education or motivation to pursue an academic career so that defiers can be ruled out.

Monotonicity has been considered in Lee (2009), Zhang and Rubin (2003), and Zhang *et al.* (2008) to bound the ATE on the always observed (stratum 11) and in Lechner and

Melly (2007) to derive bounds for the treated and observed population. Lee (2009) shows that the following bounds are sharp for the ATE on the always observed:

$$\begin{aligned}\Delta_{11}^{\mathrm{UB}} &= \bar{Y}_{1,1}(\max|P_{1|0}/P_{1|1}) - \bar{Y}_{0,1},\\ \Delta_{11}^{\mathrm{LB}} &= \bar{Y}_{1,1}(\min|P_{1|0}/P_{1|1}) - \bar{Y}_{0,1}.\end{aligned} \tag{11}$$

Under monotonicity, $o(0,1)$ consists only of individuals belonging to stratum 11 so that $\bar{Y}_{0,1}$ is the mean potential outcome of the always observed under non-treatment. Furthermore, $P_{1|0} = \pi_{11}$. Therefore, the share of the always observed in $o(1,1)$ is $\pi_{11}/(\pi_{11} + \pi_{10}) = P_{1|0}/P_{1|1}$. In the most extreme cases, either the upper or lower $P_{1|0}/P_{1|1}$ share of the outcome distribution in $o(1,1)$ represents the potential outcomes of the always observed under treatment, which gives rise to the upper and lower bounds on Δ_{11} that are tighter than the worst case bounds.

Two points are worth noting. First, we have seen in the last section that if $\pi_{00} > \pi_{11}$, informative bounds are only obtained for the observed population and either compliers or defiers without further assumptions. Introducing monotonicity also identifies informative bounds for the always observed, which turn out to be tighter than under the stochastic dominance assumption discussed below. Second, if $P_{1|0} - P_{1|1} > 0$, the bounds are not informative, because π_{01} cannot be zero. As discussed before, the data can provide evidence against (positive or negative) monotonicity.

We now derive the bounds on the ATE on the compliers, Δ_{10}, which are just special cases of the worst case bounds given that $\pi_{01} = 0$. Therefore, they are sharp given the sharpness of the worst case bounds. Thus, under monotonicity Δ_{10} is bounded by

$$\begin{aligned}\Delta_{10}^{\mathrm{UB}} &= \bar{Y}_{1,1}(\max|(P_{1|1} - P_{1|0})/P_{1|1}) - Y^{\mathrm{LB}},\\ \Delta_{10}^{\mathrm{LB}} &= \bar{Y}_{1,1}(\min|(P_{1|1} - P_{1|0})/P_{1|1}) - Y^{\mathrm{UB}}.\end{aligned} \tag{12}$$

Monotonicity does not shrink the bounds for the compliers, as the worst case bounds under non-treatment are unaffected by ruling out defiers. However, the assumption assures that the bounds are informative. Indeed, in the worst case scenario the bounds were only informative if $P_{1|0} - P_{1|1} < 0$ which implies that the lower bound on the defiers' share is zero ($\pi_{01}^{\min} = 0$), see equation (4).

Assumption 3 has identifying power for the observed population, which is now only a mixture of always observed and compliers. The respective proportions of these groups are

$$\begin{aligned}\frac{\pi_{11}}{\pi_{11} + \Pr(T=1)\cdot\pi_{10}} &= \frac{P_{1|0}}{\Pr(S=1)},\\ \frac{\Pr(T=1)\cdot\pi_{10}}{\pi_{11} + \Pr(T=1)\cdot\pi_{10}} &= \frac{\Pr(T=1)\cdot(P_{1|1} - P_{1|0})}{\Pr(S=1)}.\end{aligned}$$

Again, the bounds are a special case of the worst case bounds under $\pi_{01} = 0$ and given by

$$\begin{aligned}\Delta_{S=1}^{\mathrm{UB}} &= \frac{P_{1|0}}{\Pr(S=1)} \cdot (\Pr(T=0) \cdot \bar{Y}_{1,1}(\max|P_{1|0}/P_{1|1}) - \bar{Y}_{0,1}) \\ &\quad + \frac{\Pr(T=1) \cdot P_{1|1}}{\Pr(S=1)} \cdot \bar{Y}_{1,1} - \frac{\Pr(T=1) \cdot (P_{1|1} - P_{1|0})}{\Pr(S=1)} \cdot Y^{\mathrm{LB}},\end{aligned} \tag{13}$$

$$\begin{aligned}\Delta_{S=1}^{\mathrm{UB}} &= \frac{P_{1|0}}{\Pr(S=1)} \cdot (\Pr(T=0) \cdot \bar{Y}_{1,1}(\min|P_{1|0}/P_{1|1}) - \bar{Y}_{0,1}) \\ &\quad + \frac{\Pr(T=1) \cdot P_{1|1}}{\Pr(S=1)} \cdot \bar{Y}_{1,1} - \frac{\Pr(T=1) \cdot (P_{1|1} - P_{1|0})}{\Pr(S=1)} \cdot Y^{\mathrm{UB}},\end{aligned} \tag{14}$$

The identification region shrinks as the complier population decreases and $\Delta_{S=1}$ is point identified in the absence of compliers so that $P_{1|1} - P_{1|0} = 0$. Then, the observed population consists only of always observed individuals.

Stochastic dominance

Assumption 4 formalizes stochastic dominance which has been considered by Zhang and Rubin (2003), Grilli and Mealli (2008), Zhang *et al.* (2008), and Lechner and Melly (2007), see also Blundell *et al.* (2007) for a related, but somewhat different form of dominance.

Assumption 4. $\Pr(Y_i(t) \leqslant y|G = 11) \leqslant \Pr(Y_i(t) \leqslant y|G = 10), \quad \forall y \in [Y^{\mathrm{LB}}, Y^{\mathrm{UB}}]$, $t \in \{0,1\}$, and

$\Pr(Y_i(t) \leqslant y|G = 11) \leqslant \Pr(Y_i(t) \leqslant y|G = 01), \quad \forall y \in [Y^{\mathrm{LB}}, Y^{\mathrm{UB}}], t \in \{0,1\}$ (stochastic dominance).

That is, the potential outcome among the always observed at any rank of the outcome distribution and in any treatment state is at least as high as that of the compliers or the defiers, respectively.[5] Taking the evaluation of the returns to a job training as example, it implies that the always observed have potential wages that are at least as high as the ones of other groups. To justify Assumption 4, note that the always observed are employed irrespective of the training. Therefore, they are likely to be more motivated and/or able than other populations. Zhang *et al.* (2008) argue that ability tends to be positively correlated with wages and thus, the stochastic dominance assumption (or 'positive selection') appears to be plausible. Similar arguments hold for the evaluation of private schooling with regard to the performance in college entrance examinations. As the always observed are those taking the exam with and without private schooling, it seems reasonable to assume that their potential test scores are higher than those of other groups.

Under Assumption 4, Imai (2008) shows that the following bounds proposed by Zhang and Rubin (2003) are sharp for the ATE on the always observed:

$$\begin{aligned}\Delta_{11}^{\mathrm{UB}} &= \bar{Y}_{1,1}(\max|(P_{1|0} - \pi_{01}^{\max})/P_{1|1}) - \bar{Y}_{0,1}, \\ \Delta_{11}^{\mathrm{LB}} &= \bar{Y}_{1,1} - \bar{Y}_{0,1}(\max|(P_{1|0} - \pi_{01}^{\max})/P_{1|0}).\end{aligned} \tag{15}$$

As $E[Y(t)|G = 11] \geqslant E[Y(t)|G = 10]$, $E[Y(t)|G = 11] \geqslant E[Y(t)|G = 01]$ for $t \in \{0,1\}$, the means $\bar{Y}_{1,1}$, $\bar{Y}_{0,1}$ constitute the lower bounds of $E[Y(1)|G = 11]$ and $E[Y(0)|G = 11]$,

[5] For our purpose, which is the derivation of bounds on the ATE, the weaker mean dominance assumption, that is, $E[Y(t)|G = 11] \geqslant E[Y(t)|G = 10]$ and $E[Y(t)|G = 11] \geqslant E[Y(t)|G = 01], t \in \{0,1\}$, is sufficient. However, stochastic dominance is required when considering other parameters as for instance the quantile treatment effect.

respectively. Thus, Assumption 4 is likely to shrink the worst case bounds because $\bar{Y}_{0,1} \geqslant \bar{Y}_{0,1}(\min|(P_{1|0} - \pi_{01}^{\max})/P_{1|0})$ and $\bar{Y}_{1,1} \geqslant \bar{Y}_{1,1}(\min|(P_{1|0} - \pi_{01}^{\max})/P_{1|1})$. Note that width of the bounds is maximized if the share of the always observed is smaller than the one of the never observed. Then, $\bar{Y}_{1,1}(\max|(P_{1|0} - \pi_{01}^{\max})/P_{1|1}) = \bar{Y}_{1,1}(\max|0)$ and $\bar{Y}_{0,1}(\max|(P_{1|0} - \pi_{01}^{\max})/P_{1|0}) = \bar{Y}_{0,1}(\max|0)$ so that they are uninformative which requires us to use the theoretical upper bound Y^{UB}.

Stochastic dominance implies the following bounds for the ATE on the compliers:

$$\begin{aligned}\Delta_{10}^{\text{UB}} &= \bar{Y}_{1,1} - Y^{\text{LB}},\\ \Delta_{10}^{\text{LB}} &= \min_{\pi_{01}} \left[\bar{Y}_{1,1}(\min|(P_{1|1} - P_{1|0} + \pi_{01})/P_{1|1}) - \bar{Y}_{0,1}(\max|(P_{1|0} - \pi_{01})/P_{1|0})\right].\end{aligned} \tag{16}$$

The intuition is that any mean potential outcome of the compliers is at best as high as that of the always observed, so that $\bar{Y}_{1,1}$ and $\bar{Y}_{0,1}(\max|(P_{1|0} - \pi_{01})/P_{1|0})$ are upper bounds for $E[Y(1)|G = 10]$ and $E[Y(0)|G = 10]$, respectively. Thus, the bounds are likely tighter than the worst case bounds as $\bar{Y}_{1,1} \leqslant \bar{Y}_{1,1}(\max|(P_{1|1} - P_{1|0} + \pi_{01})/P_{1|1})$ and $\bar{Y}_{0,1}(\max|(P_{1|0} - \pi_{01})/P_{1|0}) \leqslant Y^{\text{UB}}$. In particular, stochastic dominance in general raises the lower bound of the effect, as it does not depend on Y^{UB} anymore. This is relevant for empirical applications, where the lower bound is often more interesting than the upper bound, as it provides evidence on the existence of a positive effect. Note that, as $\bar{Y}_{1,1}(\min|(P_{1|1} - P_{1|0} + \pi_{01})/P_{1|1})$ is minimized for $\pi_{01} = \pi_{01}^{\min}$ and $\bar{Y}_{0,1}(\max|(P_{1|0} - \pi_{01})/P_{1|0})$ is maximized for $\pi_{01} = \pi_{01}^{\max}$, we need to minimize Δ_{10}^{LB} over all possible values of π_{01}.

In an analogous way, the bounds of the ATE on the defiers can be derived as

$$\begin{aligned}\Delta_{01}^{\text{UB}} &= \max_{\pi_{01}} \left[\bar{Y}_{1,1}(\max|(P_{1|0} - \pi_{01})/P_{1|1}) - \bar{Y}_{0,1}(\min|\pi_{01}/P_{1|0})\right],\\ \Delta_{01}^{\text{LB}} &= Y^{\text{LB}} - \bar{Y}_{0,1}.\end{aligned} \tag{17}$$

As for the compliers, any mean potential outcome of the defiers can be at best as high as the one of the always observed so that $\bar{Y}_{1,1}(\max|(P_{1|0} - \pi_{01})/P_{1|1})$ constitutes the upper bound under treatment and $\bar{Y}_{0,1}$ the upper bound under non-treatment. These bounds are likely to be narrower than the worst case bounds as $\bar{Y}_{1,1}(\max|(P_{1|0} - \pi_{01})/P_{1|1}) \leqslant Y^{\text{UB}}$ and $\bar{Y}_{0,1} \leqslant \bar{Y}_{0,1}(\max|\pi_{01}/P_{1|0})$. As $\bar{Y}_{1,1}(\max|(P_{1|0} - \pi_{01})/P_{1|1})$ is maximized for $\pi_{01} = \pi_{01}^{\max}$ and $\bar{Y}_{0,1}(\min|\pi_{01}/P_{1|0})$ is minimized for $\pi_{01} = \pi_{01}^{\min}$, we need to maximize Δ_{01}^{UB} over all possible values of π_{01}.

Finally, the bounds of the ATE on the observed population are identified by

$$\begin{aligned}\Delta_{S=1}^{\text{UB}} &= \frac{\Pr(T=0)\cdot P_{1|0}}{\Pr(S=1)} \cdot \bar{Y}_{1,1}(\max|(P_{1|0} - \pi_{01}^{\max})/P_{1|1})\\ &\quad - \frac{P_{1|0} - \Pr(T=1)\cdot \pi_{01}^{\max}}{\Pr(S=1)} \bar{Y}_{0,1}\\ &\quad + \frac{\Pr(T=1)\cdot P_{1|1}}{\Pr(S=1)} \cdot \bar{Y}_{1,1} - \frac{\Pr(T=1)\cdot(P_{1|1} - P_{1|0} + \pi_{01}^{\max})}{\Pr(S=1)} \cdot Y^{\text{LB}}\\ \Delta_{S=1}^{\text{LB}} &= \frac{\Pr(T=1)\cdot P_{1|1} + \Pr(T=0)\cdot(P_{1|0} - \pi_{01}^{\max})}{\Pr(S=1)} \cdot \bar{Y}_{1,1}\\ &\quad - \frac{\Pr(T=1)\cdot P_{1|1}}{\Pr(S=1)} \cdot \bar{Y}_{0,1}(\max|(P_{1|0} - \pi_{01}^{\max})/P_{1|0})\end{aligned} \tag{18}$$

$$\frac{+\Pr(T=0)\cdot\pi_{01}^{\max}}{\Pr(S=1)}\cdot Y^{\mathrm{LB}}-\frac{\Pr(T=0)\cdot P_{1|0}}{\Pr(S=1)}\cdot\bar{Y}_{0,1}. \tag{18}$$

For both the upper and the lower bound of $\Delta_{S=1}$, stochastic dominance eliminates Y^{UB} present in the worst case scenario. The identification region shrinks as the bounds for always observed, compliers and defiers become narrower. However, if the never observed outnumber the always observed, the bounds correspond to the worst case ones. Interestingly, the bounds on $\Delta_{S=1}$ are tighter than those on Δ_{11}. Again, we obtain more informative bounds for the observed population than for the always observed if $\pi_{00} > \pi_{11}$.

Monotonicity and stochastic dominance

We subsequently investigate the identifying power of combining Assumptions 3 and 4. This was first considered by Zhang and Rubin (2003) who derive the following bounds for the always observed, which were shown to be sharp by Imai (2008):

$$\begin{aligned}\Delta_{11}^{\mathrm{UB}} &= \bar{Y}_{1,1}(\max|P_{1|0}/P_{1|1})-\bar{Y}_{0,1},\\ \Delta_{11}^{\mathrm{LB}} &= \bar{Y}_{1,1}-\bar{Y}_{0,1}.\end{aligned} \tag{19}$$

These bounds are a simplification of those under stochastic dominance for $\pi_{01}=0$. The upper bound is the same as under monotonicity and is, thus, not affected by additionally assuming stochastic dominance, which does not change the conditional means to be compared. However, the lower bound is tightened by the fact that $\bar{Y}_{1,1}$ now constitutes the lower bound of the mean potential outcome of the always observed under treatment.

In the same manner, the bounds on the compliers simplify to

$$\begin{aligned}\Delta_{10}^{\mathrm{UB}} &= \bar{Y}_{1,1}-Y^{\mathrm{LB}},\\ \Delta_{10}^{\mathrm{LB}} &= \bar{Y}_{1,1}(\min|1-P_{1|0}/P_{1|1})-\bar{Y}_{0,1}.\end{aligned} \tag{20}$$

The upper bound is the same as under stochastic dominance and unaffected by adding monotonicity, because ruling out defiers does not change the comparison outcome under non-treatment, which is still the theoretical lower bound (as compliers are not observed under non-treatment). Also for the lower bound, monotonicity does not bring any benefits for the same reasons as under Assumption 3: for all admissible values $\pi_{01}\geqslant 0$, $\pi_{01}=0$ minimizes the lower bound of the mean potential outcome under treatment. Therefore, setting $\pi_{01}=0$ by assumption does neither increase the lower bound of the mean potential outcome, nor of Δ_{10}.

The bounds of the ATE on the observed population are identified by

$$\begin{aligned}\Delta_{S=1}^{\mathrm{UB}} &= \frac{P_{1|0}}{\Pr(S=1)}\cdot(\Pr(T=0)\cdot\bar{Y}_{1,1}(\max|P_{1|0}/P_{1|1})-\bar{Y}_{0,1})\\ &\quad+\frac{\Pr(T=1)\cdot P_{1|1}}{\Pr(S=1)}\cdot\bar{Y}_{1,1}-\frac{\Pr(T=1)\cdot(P_{1|1}-P_{1|0})}{\Pr(S=1)}\cdot Y^{\mathrm{LB}},\end{aligned}$$

$$\begin{aligned}\Delta_{S=1}^{\mathrm{LB}} &= \frac{P_{1|0}}{\Pr(S=1)}\cdot(\bar{Y}_{1,1}-\bar{Y}_{0,1})+\frac{\Pr(T=1)\cdot(P_{1|1}-P_{1|0})}{\Pr(S=1)}\cdot(\bar{Y}_{1,1}-\bar{Y}_{0,1})\\ &= \bar{Y}_{1,1}-\bar{Y}_{0,1}.\end{aligned} \tag{21}$$

Compared to just invoking monotonicity, the upper bound of $\Delta_{S=1}$ is unaffected by the introduction of stochastic dominance. This is due to the fact that $\bar{Y}_{1,1}$ still represents the weighted average of the mean potential outcomes under treatment of the always observed and the compliers (even if the potential outcomes are now restricted in a particular way by stochastic dominance). Nor does the assumption change the bound of any other potential outcome relevant to the upper bound. Stochastic dominance does, however, change the lower bound on $\Delta_{S=1}$. $\bar{Y}_{0,1}$ now represents the mean potential outcome under non-treatment for all observed individuals because it constitutes the upper bound on the compliers' mean potential outcome. Therefore, an interesting result of imposing both assumptions is that the lower bound now coincides with the one for the always observed.

IV. Estimation

This section briefly sketches estimation, which is mostly based on the sample analogs of the bounds derived under the various assumptions (even though the lower bound for the compliers and the upper bound for the defiers under stochastic dominance deserve particular consideration, as discussed below). To this end, we define the following sample parameters:

$$\hat{P}_{1|1} \equiv \frac{\sum_{i=1}^{n} S_i \cdot T_i}{\sum_{i=1}^{n} T_i}, \qquad \hat{P}_{0|1} \equiv 1 - \frac{\sum_{i=1}^{n} S_i \cdot T_i}{\sum_{i=1}^{n} T_i},$$

$$\hat{P}_{1|0} \equiv \frac{\sum_{i=1}^{n} S_i \cdot (1 - T_i)}{\sum_{i=1}^{n} (1 - T_i)}, \qquad \hat{P}_{0|0} \equiv 1 - \frac{\sum_{i=1}^{n} S_i \cdot (1 - T_i)}{\sum_{i=1}^{n} (1 - T_i)},$$

$$\hat{\bar{Y}}_{1,1} \equiv \frac{\sum_{i=1}^{n} Y_i \cdot S_i \cdot T_i}{\sum_{i=1}^{n} S_i \cdot T_i}, \qquad \hat{\bar{Y}}_{0,1} \equiv \frac{\sum_{i=1}^{n} Y_i \cdot S_i \cdot (1 - T_i)}{\sum_{i=1}^{n} S_i \cdot (1 - T_i)},$$

$$\hat{\bar{Y}}_{t,s}(\max|q) \equiv \frac{\sum_{i=1}^{n} Y_i \cdot I\{S_i = s\} \cdot I\{T_i = t\} \cdot I\{Y \geqslant \hat{y}_{1-q}\}}{\sum_{i=1}^{n} I\{S_i = s\} \cdot I\{T_i = t\} \cdot I\{Y \geqslant \hat{y}_{1-q}\}},$$

$$\hat{\bar{Y}}_{t,s}(\min|q) \equiv \frac{\sum_{i=1}^{n} Y_i \cdot I\{S_i = s\} \cdot I\{T_i = t\} \cdot I\{Y \leqslant \hat{y}_{q}\}}{\sum_{i=1}^{n} I\{S_i = s\} \cdot I\{T_i = t\} \cdot I\{Y \leqslant \hat{y}_{q}\}},$$

$$\hat{y}_q \equiv \min\left\{ y : \frac{\sum_{i=1}^{n} S_i \cdot T_i \cdot I\{Y_i \leqslant y\}}{\sum_{i=1}^{n} S_i \cdot T_i} \geqslant q \right\},$$

where $I\{\cdot\}$ is the indicator function. Using these expressions instead of the population parameters in the formulas for the bounds immediately yields feasible estimators. However, note that depending on the parameters considered, particular common support conditions have to be satisfied. For example, the estimation of $\hat{P}_{1|1}, \hat{P}_{0|1}$ and $\hat{P}_{1|0}, \hat{P}_{0|0}$ requires that $\Pr(T = 1) > 0$ and $\Pr(T = 1) < 1$, respectively (or that $0 < \Pr(T = 1) < 1$ for the joint estimation of $\hat{P}_{1|1}, \hat{P}_{0|1}, \hat{P}_{1|0}, \hat{P}_{0|0}$). Likewise, $\hat{\bar{Y}}_{1,1}$ demands that $E(S \cdot D) > 0$ and $\hat{\bar{Y}}_{0,1}$ that $E(S \cdot D) < 1$.

$\sqrt{n}$-consistency and asymptotic normality of the estimators of the bounds for the compliers and the observed population under both monotonicity and stochastic dominance directly follows from the results of Lee (2009). To see this, first consider the estimators of $\Delta_{11}^{\mathrm{UB}}, \Delta_{11}^{\mathrm{LB}}$ under monotonicity alone:

$$\hat{\Delta}_{11}^{\mathrm{UB}} = \hat{\bar{Y}}_{1,1}(\max|\hat{P}_{1|0}/\hat{P}_{1|1}) - \hat{\bar{Y}}_{0,1},$$
$$\hat{\Delta}_{11}^{\mathrm{LB}} = \hat{\bar{Y}}_{1,1}(\min|\hat{P}_{1|0}/\hat{P}_{1|1}) - \hat{\bar{Y}}_{0,1}.$$

In his appendix, Lee (2009) shows $\sqrt{n}$-consistency and asymptotic normality using a GMM framework based on theorems 2.6 and 7.2 of Newey and McFadden (1994). It suffices to show the desirable properties for $\hat{\bar{Y}}_{1,1}(\max|\hat{P}_{1|0}/\hat{P}_{1|1})$ and $\hat{\bar{Y}}_{1,1}(\max|\hat{P}_{1|0}/\hat{P}_{1|1})$ (or just one of them due to the symmetry of the problem) because these estimators are independent of the observed mean outcome under non-treatment $\hat{\bar{Y}}_{0,1}$.

Now consider the estimators for the compliers under monotonicity:

$$\begin{aligned}\hat{\Delta}_{10}^{\mathrm{UB}} &= \hat{\bar{Y}}_{1,1}(\max|1-P_{1|0}/P_{1|1}) - Y^{\mathrm{LB}},\\ \hat{\Delta}_{10}^{\mathrm{LB}} &= \hat{\bar{Y}}_{1,1}(\min|1-P_{1|0}/P_{1|1}) - Y^{\mathrm{UB}}.\end{aligned} \tag{22}$$

$Y^{\mathrm{LB}}, Y^{\mathrm{UB}}$ are constants not relevant for the properties of the estimators. Furthermore, note that the problem of estimating $\hat{\bar{Y}}_{1,1}(\max|1 - P_{1|0}/P_{1|1})$ is symmetric to $\hat{\bar{Y}}_{1,1}(\max|\hat{P}_{1|0}/\hat{P}_{1|1})$ (and $\hat{\bar{Y}}_{1,1}(\min|1 - P_{1|0}/P_{1|1})$ to $\hat{\bar{Y}}_{1,1}(\max|\hat{P}_{1|0}/\hat{P}_{1|1})$). Therefore, Lee's results immediately apply to the estimators of the bounds for the compliers. This in turn implies $\sqrt{n}$-consistency and asymptotic normality of $\hat{\Delta}_{S=1}^{\mathrm{UB}}, \hat{\Delta}_{S=1}^{\mathrm{LB}}$, as the observed population is just a weighted average of the always observed and compliers. Finally, note that imposing stochastic dominance in addition to monotonicity replaces some parameters in the estimators by simple conditional means, which again entails $\sqrt{n}$-consistency and asymptotic normality of all estimators.

However, under stochastic dominance alone, the latter result does not apply to the lower bound of the ATE on the compliers and the upper bound of the ATE on the defiers, because they contain min and max operators, respectively. Hirano and Porter (2012) show that for parameters that are non-differentiable functionals of the data (such as min/max operators), asymptotically unbiased estimators do not exist. Therefore, the sample analog estimators of $\Delta_{10}^{\mathrm{LB}} = \min_{\pi_{01}}\left[\bar{Y}_{1,1}(\min|(P_{1|1} - P_{1|0} + \pi_{01})/P_{1|1}) - \bar{Y}_{0,1}(\max|(P_{1|0} - \pi_{01})/P_{1|0})\right]$ and $\Delta_{01}^{\mathrm{UB}} = \max_{\pi_{01}}\left[\bar{Y}_{1,1}(\max|(P_{1|0} - \pi_{01})/P_{1|1}) - \bar{Y}_{0,1}(\min|\pi_{01}/P_{1|0})\right]$ are likely downward and upward biased, respectively, due to optimizing over the defier proportion. This yields overly conservative (i.e. too large) intervals for the ATEs as well as confidence regions that are based on standard asymptotics or bootstrapping. In our application (see the next section), we in addition to bootstrap-based inference also consider the method of Chernozhukov *et al.* (2013) to obtain half-median-unbiased point estimates and confidence intervals for the lower bound of the compliers. The procedure is described in the Supporting Information.

V. Application

In this section, we use our methods to re-evaluate the school voucher experiment of Angrist *et al.* (2006). As mentioned before, the authors investigate the effects of school vouchers provided to high school students in the course of Colombia's Programa de Amplicación de Cobertura de la Educación Secundaria (PACES) program (taking place between 1991 and 1997). The outcome we focus on are the reading scores achieved in the centralized

college entrance examinations, the ICFES, several years later. Many of the vouchers that covered half the cost of private secondary schooling were randomly assigned by a lottery among applicants so that Assumption 2 appears likely to hold. The experimental estimates in Angrist *et al.* (2006) suggest that vouchers increase reading test scores on average by roughly 0.7 points (or roughly 0.12 standard deviations) and this result is significant at the 5% level.

However, only 30.2% (or 1,223 students) of the 4,044 applicants actually took the test. Therefore, the experimental estimates might be flawed by selection bias. For example, if the treatment positively affects the likelihood to take the test so that also *a priori* less motivated students are induced to participate, then the distribution of motivation differs across treated and non-treated students conditional on being tested. If motivation positively affects the test scores, this entails a (downward) bias of the estimated effect. For this reason, Angrist *et al.* (2006) use both censored regression to control for sample selection and derive non-parametric bounds on the ATE of the always observed population based on Assumptions 3 (monotonicity of selection) and 5 (monotone treatment response). On balance, they still find substantial gains from the PACES program.

We complement their analysis by estimating the ATE under different sets of assumptions and for several populations. To be specific, we invoke Assumption 3 (monotonicity of selection) and/or Assumption 4 (stochastic dominance) to bound the ATE on the always observed, compliers and the observed population. Both assumptions appear to be plausible in this context. Monotonicity roots in the presumption that the treatment weakly increases participation in the exam because private schools are plausibly more committed to the academic success of their (paying) students, which may serve as measure of school quality. Stochastic dominance seems reasonable because the always observed are those taking the exam irrespective of the treatment and are, thus, likely to have higher potential test scores than other groups, for instance due to ability or motivation. We do not consider Assumption 5 (monotone treatment response) which restricts the direction of the effects.

Estimation is based on the approach outlined in the Supporting Information. Concerning inference, we compute the confidence intervals based on the method described in Imbens and Manski (2004), which contains the treatment effect of interest with a probability of at least 95%:

$$(\hat{\Delta}^{\mathrm{LB}} - 1.645 \cdot \hat{\sigma}^{\mathrm{LB}}, \hat{\Delta}^{\mathrm{UB}} + 1.645 \cdot \hat{\sigma}^{\mathrm{UB}}),$$

where $\hat{\Delta}^{\mathrm{LB}}, \hat{\Delta}^{\mathrm{UB}}$ are the estimated bounds and $\hat{\sigma}^{\mathrm{LB}}, \hat{\sigma}^{\mathrm{UB}}$ denote their respective estimated standard errors.[6] We compute the latter by bootstrapping the original sample 1,999 times and estimating $\hat{\Delta}^{\mathrm{LB}}, \hat{\Delta}^{\mathrm{UB}}$ in each bootstrap replication in order to estimate their distributions. As worst case bounds Y^{UB} and Y^{LB}, we take the maximum and minimum test scores observed among test takers.

The estimates of the conditional selection probabilities, $\hat{P}_{1|1} = 0.328$, $\hat{P}_{1|0} = 0.267$, $\hat{P}_{0|1} = 0.672$ and $\hat{P}_{0|0} = 0.733$, allow us to bound the strata proportions. Table 4 reports these bounds and shows that the lower bound on the share of the never observed is larger than the upper bound on the share of any other population and in particular than the one of

[6] The confidence intervals apply to cases where the distance between the upper and lower bound of the effect is bounded away from zero, see the discussion in Stoye (2009).

TABLE 4

Estimated (bounds on the) proportions of latent strata

Latent strata	*Bounds without monotonicity*	*Proportions under monotonicity*
Always observed	[0.000, 0.267]	0.267
Compliers	[0.061, 0.328]	0.061
Never observed	[0.406, 0.672]	0.672
Defiers	[0.000, 0.267]	—
Always observed among observed		0.897
Compliers among observed		0.103

the always observed. Therefore, without monotonicity the bounds on this population will be uninformative in the worst case scenario and quite large under stochastic dominance. Moreover, the lower bound of the compliers' share is larger than zero so that positive monotonicity is consistent with the data whereas negative is not. In Table 4, we also provide the estimated strata proportions and the mixture probabilities under Assumption 3 (monotonicity), which are then point identified.

Table 5 presents the results for the always observed, compliers and the observed population under various assumptions. The bounds of the ATE estimates are given in square brackets, the 95% confidence intervals in round brackets. The worst case bounds are not informative for the always observed and very wide for any other population. Monotonicity narrows the bounds substantially for the always observed and the observed population, even though the identification region still includes the zero. As discussed before, monotonicity has no identifying power for the compliers as a zero proportion of π_{01} implies the widest bounds possible.

Stochastic dominance entails narrower bounds than the worst case scenario for all three populations. However, for the always observed, the identification region is substantially larger than under monotonicity. Using both assumptions jointly brings important improvements. The lower bounds of the ATEs on the always observed and the observed population are now significantly larger than zero and point to a positive effect of private schooling. Also, the upper bounds do not appear unreasonably high. For the observed population, this is due to the small share of compliers (10.28%) to which the theoretical upper bound Y^{UB} applies. For the compliers alone, the bounds are not more informative than under stochastic dominance, as monotonicity does not further narrow the bounds for reasons discussed in section III.

All in all our results give support to the conclusions of Angrist *et al.* (2006) suggesting that the PACES program in Colombia had a positive effect on the reading scores in college entrance examinations. The lower bounds of the ATEs on those who would take the test irrespective of private schooling (supposedly the most able and motivated) and on all test takers are positive when invoking both monotonicity and stochastic dominance. Furthermore, the Imbens and Manski (2004) confidence intervals suggest that these ATEs are significantly different from zero. For the compliers alone, however, we cannot reject the null hypothesis of a zero effect based on our assumptions.

TABLE 5

ATE estimates and confidence intervals

Assumptions	*Always observed*	*Compliers*	*Observed*
Worst case	[−31.000, 32.000]	[−24.593, 25.432]	[−16.587, 17.413]
	Not informative	(−26.503, 27.086)	(−17.483, 18.276)
Monotonicity	[−1.113, 2.547]	[−24.593, 25.432]	[−7.736, 8.645]
	(−1.892, 3.308)	(−26.503, 27.086)	(−8.529, 9.450)
Stochastic dominance	[−13.396, 17.079]	[−13.396, 17.604]	[−14.676, 17.413]
	(−15.064, 18.344)	(−15.042, 18.754)	(−15.856, 18.276)
Monotonicity + stochastic dominance	[0.683, 2.547]	[−7.514, 17.604]	[0.683, 3.369]
	(0.140, 3.308)	(−9.132, 18.754)	(0.140, 4.423)

Notes: Bounds in square brackets and confidence intervals in round brackets. Confidence intervals are based on 1,999 bootstraps. All results are based on the estimators for discrete outcomes discussed in the Supporting Information. Applying the procedure of Chernozhukov *et al.* (2013) to the lower bound of the ATE on the compliers under stochastic dominance gives an estimate of −12.428 with a lower confidence bound of −13.796. In the algorithm described in the Supporting Information, we set $\alpha = 0.05$, $m = 100$, $B = 1,999$ and $R = 200,000$. We thank Xuan Chen and Carlos Flores for providing us with the Matlab code of their study Chen and Flores (2012), which implements the Chernozhukov *et al.* (2013) procedure and for their helpful advice about its use.

VI. Conclusion

This article discusses the partial identification of ATE in the presence of sample selection, implying that outcomes are only observed for a non-random subpopulation. The previous work considering this problem has predominantly focussed on bounding the ATE on the 'always observed', whose outcomes are observed irrespective of the treatment received. Here, we also derived sharp bounds for other populations such as the 'compliers' (observed under treatment, not observed under non-treatment) and the observed population (all individuals whose outcomes are observed), which is a mixture of several groups.

These populations appear to be relevant for policy recommendations in many empirical contexts. Taking, for instance, the compliers, one might be interested whether switching the selection state as a reaction on the treatment comes along with (and may be rooted in) a particular treatment effect. An example is the effect of a training on wages, which might induce formerly unemployed individuals to work because their potential wage surpasses their reservation wage after the training. Furthermore, it might be preferable to make causal statements rather for larger than for smaller shares of the total population. The largest subgroup for which outcomes are observed is the observed population, so that results obtained for these individuals are likely to have more external validity than those based on smaller (and unobservable) subgroups.

In the discussion on identification, we have argued that the combination of monotonicity (of selection in the treatment) and stochastic dominance (of the potential outcomes of the always observed over those of others) assumptions may bear considerable identifying power even for populations whose outcomes are, in contrast to the always observed, only observed in one treatment state. In particular, it has been shown that the lower bound of the ATE on the observed population coincides with the lower bound for the always observed. This is an important result, as we are often most interested in the lower bound, which gives evidence about the existence of a positive effect. Its practical relevance has been demonstrated by means of an empirical application to a school voucher experiment.

Finally, the article also shows that principal stratification provides an adequate framework for a better understanding of the identifying assumptions involved, because they are expressed in terms of individual selection behaviour rather than the less tangible relation of error terms in some structural model. For example, we have found that if the share of the always observed is smaller than the one of never observed, bounds on the always observed are not informative if we do not assume monotonicity of selection in the treatment. In contrast, we can still bound the ATE on the observed population. This might be hard to see from the equations characterizing a structural model.

References

Angrist, J., Bettinger, E. and Kremer, M. (2006). 'Long-term educational consequences of secondary school vouchers: evidence from administrative records in Colombia', *American Economic Review*, Vol. 96, pp. 847–862.

Angrist, J., Imbens, G. and Rubin, D. (1996). 'Identification of causal effects using instrumental variables', *Journal of American Statistical Association*, Vol. 91, pp. 444–472.

Blanco, G., Flores, C. A. and Flores-Lagunes, A. (2011). *Bounds on Quantile Treatment Effects of Job Corps on Participants' Wages*, IZA Discussion Paper No. 6065.

Blundell, R., Gosling, A., Ichimura, H. and Meghir, C. (2007). 'Changes in the distribution of male and female wages accounting for employment composition using bounds', *Econometrica*, Vol. 75, pp. 323–363.

Chen, X. and Flores, C. A. (2012). Bounds on treatment effects in the presence of sample selection and noncompliance: the wage effects of Job Corps, Working Paper, Department of Economics, University of Miami, Miami.

Chernozhukov, V., Lee, S. and Rosen, A. (2013). 'Intersection bounds: estimation and inference', *Econometrica*, Vol. 81, pp. 667–737.

Das, M., Newey, W. K. and Vella, F. (2003). 'Nonparametric estimation of sample selection models', *Review of Economic Studies*, Vol. 70, pp. 33–58.

Fisher, R. (1935). *The Design of Experiments*, Oliver and Boyd, Edinburgh.

Frangakis, C. E. and Rubin, D. B. (2002). 'Principal stratification in causal inference', *Biometrics*, Vol. 58, pp. 21–29.

Grilli, L. and Mealli, F. (2008). 'Nonparametric bounds on the causal effect of university studies on job opportunities using principal stratification', *Journal of Educational and Behavioral Statistics*, Vol. 33, pp. 111–130.

Gronau, R. (1974). 'Wage comparisons – a selectivity bias', *Journal of Political Economy*, Vol. 82, pp. 1119–1143.

Hausman, J. A. and Wise, D. A. (1979). 'Attrition bias in experimental and panel data: the Gary income maintenance experiment', *Econometrica*, Vol. 47, pp. 455–473.

Heckman, J. J. (1974). 'Shadow prices, market wages and labor supply', *Econometrica*, Vol. 42, pp. 679–694.

Heckman, J. 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*, Vol. 5, pp. 475–492.

Heckman, J. J. (1979). 'Sample selection bias as a specification error', *Econometrica*, Vol. 47, pp. 153–161.

Hirano, K. and Porter, J. R. (2012). 'Impossibility results for nondifferentiable functionals', *Econometrica*, Vol. 80, pp. 1769–1790.

Horowitz, J. and Manski, C. F. (1995). 'Identification and robustness with contaminated and corrupted data', *Econometrica*, Vol. 63, pp. 281–302.

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

Horowitz, J. and Manski, C. F. (2000). 'Nonparametric analysis of randomized experiments with missing covariate and outcome data', *Journal of the American Statistical Association*, Vol. 95, pp. 77–84.

Huber, M. (2012). 'Identification of average treatment effects in social experiments under alternative forms of attrition', *Journal of Educational and Behavioral Statistics*, Vol. 37, pp. 443–474.
Huber, M. and Mellace, G. (2013). 'Testing exclusion restrictions and additive separability in sample selection models', *Empirical Economics*, forthcoming. DOI: 10.1007/s00181-013-0742-1.
Huber, M. and Melly, B. (2012). A test of the conditional independence assumption in sample selection models, Discussion Paper No. 2012-11, Department of Economics, Brown University.
Imai, K. (2008). 'Sharp bounds on the causal effects in randomized experiments with 'truncation-by-death'', *Statistics & Probability Letters*, Vol. 78, pp. 144–149.
Imbens, G. W. (2004). 'Nonparametric estimation of average treatment effects under exogeneity: a review', *Review of Economics and Statistics*, Vol. 86, pp. 4–29.
Imbens, G. W. (2007). 'Nonadditive Models with Endogenous Regressors', in Blundell R, Newey W. and Persson T (eds), *Advances in Economic and Econometrics: Theory and Applications*, Ninth World Congress. III, Cambridge: Cambridge University Press, pp. 17–46.
Imbens, G. W. and Angrist, J. (1994). 'Identification and estimation of local average treatment effects', *Econometrica*, Vol. 62, pp. 467–475.
Imbens, G. W. and Manski, C. F. (2004). 'Confidence intervals for partially identified parameters', *Econometrica*, Vol. 72, pp. 1845–1857.
Imbens, G. W. and Wooldridge, J. (2009). 'Recent developments in the econometrics of program evaluation', *Journal of Economic Literature*, Vol. 47, pp. 5–86.
Lechner, M. and Melly, B. (2007). Earnings Effects of Training Programs, IZA Discussion Paper No. 2926, IZA, Bonn.
Lee, D. S. (2009). 'Training wages and sample selection: estimating sharp bounds on treatment effects', *Review of Economic Studies*, Vol. 76, pp. 1071–1102.
Manski, C. F. (1989). 'Anatomy of the selection problem', *Journal of Human Resources*, Vol. 24, pp. 343–360.
Manski, C. F. (1994). 'The selection problem', in Sims C. (ed.), *Advances in Econometrics: Sixth World Congress*, Cambridge: University Press, pp. 143–170.
Mealli, F. and Pacini, B. (2008). 'Comparing principal stratification and selection models in parametric causal inference with nonignorable missingness', *Computational Statistics & Data Analysis*, Vol. 53, pp. 507–516.
Mellace, C. and Rocci, R. (2011). Principal stratification in sample selection problems with non normal error terms, CEIS Research Paper No. 194, CEIS, Tor Vergata University.
Newey, W. K. (2007). 'Nonparametric continuous/discrete choice models', *International Economic Review*, Vol. 48, pp. 1429–1439.
Newey, W. K. (2009). 'Two-step series estimation of sample selection models', *Econometrics Journal*, Vol. 12, pp. 217–229.
Newey, W. K. and McFadden, D. (1994). 'Large sample estimation and hypothesis testing', in Engle R. F. and McFadden D. L. (eds), *Handbook of Econometrics*, Amsterdam: Elsevier.
Neyman, J. (1923). 'On the application of probability theory to agricultural experiments. Essay on principles', *Statistical Science*, Vol. 5, pp. 465–480.
Robins, J. (1989). 'The analysis of randomized and non-randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies', in Sechrest L., Freeman H. and Mulley A. (eds), *Health Service Research Methodology: A Focus on AIDS*, Washington DC: US Public Health Service, pp. 113–159.
Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1995). 'Analysis of semiparametric regression models for repeated outcomes in the presence of missing data', *Journal of the American Statistical Association*, Vol. 90, pp. 106–121.
Rubin, D. B. (1977). 'Assignment to treatment group on the basis of a covariate', *Journal of Educational Statistics*, Vol. 2, pp. 1–26.
Rubin, D. B. (1990). 'Formal mode of statistical inference for causal effects', *Journal of Statistical Planning and Inference*, Vol. 25, pp. 279–292.
Stoye, J. (2009). 'More on confidence intervals for partially identified parameters', *Econometrica*, Vol. 77, pp. 1299–1315.
Stoye, J. (2010). 'Partial identification of spread parameters', *Quantitative Economics*, Vol. 1, pp. 323–357.
Wooldridge, J. (2007). 'Inverse probability weighted estimation for general missing data problems', *Journal of Econometrics*, Vol. 141, pp. 1281–1301.

Zhang, J. and Rubin, D. B. (2003). 'Estimation of causal effects via principal stratification when some outcome are truncated by death', *Journal of Educational and Behavioral Statistics*, Vol. 28, pp. 353–368.
Zhang, J., Rubin, D. B. and Mealli, F. (2008). 'Evaluating the effects of job training programs on wages through principal stratification', in Millimet D., Smith J, and Vytlacil E. (eds), *Advances in Econometrics: Modelling and Evaluating Treatment Effects in Econometrics*, Amsterdam: Elsevier Science Ltd., Vol. 21, pp. 117–145.
Zhang, J., Rubin, D. B. and Mealli, F. (2009). 'Likelihood-based analysis of causal effects of job-training programs using principal stratification', *Journal of the American Statistical Association*, Vol. 104, pp. 166–176.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Appendix S1. Proofs of the identification results and inference procedures