Universität St.Gallen

Sharp bounds on causal effects under sample selection

Martin Huber and Giovanni Mellace

August 2011 Discussion Paper no. 2011-34

School of Economics and Political Science,
Department of Economics

University of St. Gallen

Editor:	Martina Flockerzi University of St. Gallen School of Economics and Political Science Department of Economics Varnbüelstrasse 19 CH-9000 St. Gallen Phone +41 71 224 23 25 Fax +41 71 224 31 35 Email seps@unisg.ch
Publisher:	School of Economics and Political Science Department of Economics University of St. Gallen Varnbüelstrasse 19 CH-9000 St. Gallen Phone +41 71 224 23 25
Electronic Publication:	Fax +41 71 224 31 35 http://www.seps.unisg.ch

Sharp bounds on causal effects under sample selection [1]

Martin Huber and Giovanni Mellace

Author's address:

Martin Huber, Ph.D.
SEW-HSG
Varnbüelstrasse 14
9000 St Gallen
Phone +41 71 224 2299
Fax +41 71 224 2302
Email Martin.Huber@unisg.ch
Website www.sew.unisg.ch

Giovanni Mellace, Ph.D.
SEW-HSG
Varnbüelstrasse 14
9000 St Gallen
Phone +41 71 224 2320
Fax +41 71 224 2302
Email Giovanni.Mellace@unisg.ch
Website www.sew.unisg.ch

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

Abstract

In many empirical problems, the evaluation of treatment effects is complicated by sample selection such 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 selected" (whose outcomes are observed irrespective of the treatment). This paper complements those studies by considering further populations, namely the "compliers" (whose selection states react to the treatment) and the selected population. We derive sharp bounds under various assumptions (monotonicity and stochastic dominance) and provide an empirical application to a school voucher experiment.

Keywords

Causal inference, principal stratification, nonparametric bounds, sample selection.

JEL Classification

C14, C21, C24.

1 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 labor, health, and educational economics. E.g., 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.

Principal stratification, see Frangakis and Rubin (2002), provides a natural framework for characterizing sample selection problems, as it allows defining populations (i.e., principal strata) in terms of their behavior w.r.t. selection under different treatment states. This is useful because the selection problem does not arise within a particular stratum consisting of individuals with the same selection behavior, i.e., being of the same "type". Thus, treatment effects are identified if the imposed assumptions and the data imply that a principal stratum is observed both under treatment and non-treatment. Therefore, the principal stratification framework enables us to explicitly state under which assumptions identification works for particular latent populations.

In the sample selection literature in economics going back to Heckman (1974, 1976, 1979), identification often relies on tight parametric restrictions, such as distributional assumptions and effect homogeneity across different populations. Albeit the literature has moved towards more flexible models, see for instance Das, Newey, and Vella (2003) and Newey (2009), it typically imposes strong assumptions on the unobserved terms likely to be violated in many applications, as confirmed by the specification test proposed in Huber and Melly (2011). In the absence of unattractive parametric restrictions and/or instruments for sample selection, treatment effects are not point identified, but upper and lower bounds can still be obtained under fairly mild restrictions. This is the approach pursued in this paper.

Partial identification of economic parameters in general goes back to Manski (1989, 1994) and Robins (1989). In the context of the sample selection problem, several contributions in the fields of principal stratification and econometrics derive bounds on treatment effects. Zhang and Rubin (2003) (see also Zhang, Rubin, and Mealli, 2008) bound the average treatment effects in one particular stratum, namely the "always selected" population, whose outcomes are observed both under treatment and non-treatment. To this end, the authors explore the identifying power of two assumptions both separately and jointly: (i) monotonicity of selection in the treatment and (ii) stochastic dominance of the potential outcomes of the always selected 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. Also Lee (2009) focuses on the always selected in the evaluation of the average wage effects of the Job Corps, the largest job training program for disadvantaged youths in the USA. As Zhang and Rubin (2003), he imposes monotonicity of selection (but not stochastic dominance) and proves sharpness of the bounds. Blanco, Flores, and Flores-Lagunes (2011) assess the same program, but add assumptions on the order of mean potential outcomes within and across subpopulations to obtain tighter bounds. Still for the always selected, Grilli and Mealli (2008) impose the stochastic dominance assumption along with a restriction on the relative size of the principal strata to bound the relative effects of two university degree programs on employment.

The main reason why the literature mainly concentrates on the always selected appears to be that it is the only principal stratum for which outcomes can be observed both under treatment and non-treatment, see the discussion in Zhang and Rubin (2003), Zhang, Rubin, and Mealli (2008), and Lee (2009). However, as argued by the latter, it depends on the context of the treatment whether this is a population of policy interest. In contrast to the aforementioned contributions, Lechner and Melly (2007) bound the effects on those treated and selected. When evaluating the wage effects of a job training program, this group corresponds to the training participants with employment and is directly observed in the data. Note that this is a mixed population consisting of always selected and those who work under treatment, but would not work without treatment, i.e., the "compliers" in selection w.r.t. treatment. Lechner and Melly (2007) argue that it is more intuitive to bound the effects for this observed group rather than a latent principal stratum. Furthermore, they claim that in the context of training programs, the treated and selected are the most interesting population because it is them who benefit from the potential wage effects of the training.

In the light of the previous discussion, we argue that it is, depending on the evaluation problem at hand, useful to look at further target populations in addition to the ones covered in the literature so far. E.g., it is the compliers whose selection state reacts on the treatment and it appears interesting in many application 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 selected. E.g., 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 selected. This might exactly be the group policy makers want to target.

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 selected population, which is again a mixture of several principal strata. Indeed, policy makers might want to learn about the average effects on all individuals whose outcomes are observed, without thinking in terms of different principal strata. Also Newey (2007) considers this population, however, investigating point identification based on continuous instruments. Finally, there is also a statistical argument to look at populations other than the always selected. In fact, if neither monotonicity nor unattractive parametric assumptions are imposed and if the share of "never selected" (whose outcomes are not observed irrespective of the treatment state) is larger than the one of the always selected, no informative bounds can be obtained for the latter. However, informative (albeit generally quite large) bounds are still available for the selected population.

The main contribution of this paper is the derivation of sharp bounds on average treatment effects among compliers, "defiers" (outcomes observed if not treated and not observed if treated), and the selected 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 strata are only observed in one treatment arm. For instance, one useful result is that under both assumptions, the lower bound on the selected population coincides with that on the always selected. 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, Bettinger, and Kremer (2006).

The remainder of this paper is organized as follows. Section 2 formally characterizes the sample selection problem based on principal stratification. Section 3 discusses partial identification of treatment effects for the compliers and the selected population under no assumptions (worst case bounds) as well as under monotonicity and/or stochastic dominance. Estimators are presented in Section 4. An empirical application to a school voucher experiment in Colombia is provided in Section 5. Section 6 concludes.

2 The selection problem

As in the standard treatment evaluation framework, assume that we are interested in the effect of a binary treatment, $T \in \{1, 0\}$, on an outcome Y, at a specific 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 $S \in \{1, 0\}$ 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:

$$\begin{aligned} S &= T \cdot S(1) + (1-T) \cdot S(0), \\ Y &= T \cdot Y(1) + (1-T) \cdot Y(0) \text{ if } S = 1 \text{ and not observed otherwise.} \end{aligned}$$

I.e., at best (if $S = 1$) 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 standard assumption 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 only conditional on some observed pre-treatment covariates X. This assumption is known in the literature as conditional independence assumption (CIA), also referred to as "selection on observables" or "unconfoundedness", see for instance Imbens (2004) and Imbens and Wooldridge (2009). It implies that the potential outcomes and selection states are independent of the treatment assignment conditional on the pre-treatment variables.

In the sample selection framework, Lee (2009), Mealli and Pacini (2008a) and Mealli and Pacini (2008b), among others, assume that the joint distribution of the potential post-treatment variables is independent of the treatment given X. Imposing joint independence is stronger than necessary. In fact, the following sequential marginal independence assumption (given X) is sufficient for identification:[1]

Assumption 2:

$T \perp (S(1), S(0)) | X = x,$

and

$(Y(1), Y(0)) \perp T | (S(1), S(0)), X = x \quad \forall x \in \mathcal{X}$ (unconfoundness),

where $\mathcal{X}$ denotes the support of X. In the further discussion, conditioning on X will be kept implicit, such that 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 principal strata (denoted as G), according to the value the selection indicator $S(t)$ takes under different treatment states. The terms "always selected", "compliers", "defiers", and "never selected" are in the spirit of Imbens and Angrist (1994) and Angrist, Imbens, and Rubin (1996). They, however, use equivalent expressions for a different problem, namely to characterize non-compliance w.r.t. instruments in the presence of endogeneity.

By Assumption 2, the stratum G some individual belongs to is independent of the treatment

[1] Assumption 2 can be replaced by $T \perp (S(1), S(0)) | X = x$ and $Y(0) \perp T | (S(1), S(0)), X = x, \quad \forall x \in \mathcal{X}$, if the inference is conditional on $T = 1$, i.e., if one is only interested in treatment effects on the treated.

Table 1: Principal strata

Principal strata (G)	S(1)	S(0)	appellation
11	1	1	Always selected
10	1	0	Compliers
01	0	1	Defiers
00	0	0	Never selected

assignment and 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 $S(1)$ or $S(0)$ but never both are known for any individual such 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 and S generate the following four observed subgroups, denoted as $o(T, S)$, which are all mixtures of two principal strata.

Table 2: Observed subgroups and principal strata

Observed subgroups $o(T, S)$	principal strata	Y 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

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:

Thus, point identification of causal effects can only be obtained by imposing unattractive parametric assumptions, see for instance the discussion in Mealli and Pacini (2008b), Zhang, Rubin, and Mealli (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

Table 3: Observed probabilities and principal strata proportions

Observed cond. selection prob.	princ. strata proportions
$P_{1\|1} \equiv \Pr(S=1\|T=1)$	$\pi_{11}+\pi_{10}$
$P_{0\|1} \equiv \Pr(S=0\|T=1)$	$\pi_{01}+\pi_{00}$
$P_{1\|0} \equiv \Pr(S=1\|T=0)$	$\pi_{11}+\pi_{01}$
$P_{0\|0} \equiv \Pr(S=0\|T=0)$	$\pi_{10}+\pi_{00}$

milder assumptions. As mentioned before, treated and non-treated units are only observed for the always selected (stratum 11), i.e. those selected 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, Rubin, and Mealli (2008), and Lee (2009), with the exception of Lechner and Melly (2007).

We, however, argue that the always selected 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 selected population ($S = 1$). Stratum 10 consists of those individuals selected with and not selected without treatment. Thus, they can be referred to as "compliers" in selection w.r.t. the treatment. This stratum is interesting in many applications as it encounters 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, consider a school voucher experiment investing the effect of private schooling on college entrance examinations. As the compliers take the test only under private schooling, they are likely to be more disadvantaged and academically less challenged at home than the always selected. This may be exactly the population policy makers want to target.

The selected population is a mixture of always selected, compliers, and defiers (stratum 01: selected under non-treatment, not selected under treatment) and therefore encounters individuals with different "selection behaviors". Still, policy makers might want to learn about the effects on all individuals whose outcomes are observed irrespective of their stratum affiliations. Also Newey (2007) considers the average effect on the selected population, however, based on a continuous instrument for selection allowing for point identification. After all, the selected population contains all subjects for which at least one outcome (under treatment and/or non-treatment) is observed such that reasonable bounds on the effects may still be attained under the assumptions discussed below. In contrast, for the never selected bounds are most likely very uninformative. Thus, the selected group appears to be the largest possible subpopulation for which useful inference appears to be feasible.

To better understand how principal stratification is related to the econometric literature on sample selection, we conclude this section by formulating the identification problem in terms of a structural model (see also Huber, 2010, Mealli and Pacini, 2008a, and Mellace and Rocci, 2010):

$$\begin{aligned} Y &= \varphi(T,U), \\ S &= I\{\varsigma(T,V) \geq 0\}, \\ T &= I\{\psi(\zeta) \geq 0\}, \end{aligned} \tag{1}$$

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 $G(V)$ such that

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

Imbens (2006) calls such a function "type of unit". Principal stratification is a natural choice of

$G(\cdot)$, as

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

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.

3 Assumptions and partial identification

The identification strategies discussed below are based on the fact that the expected values of the observed outcomes in a particular treatment state come from a mixture of two latent strata:

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

and

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

Horowitz and Manski (1995) have shown that whenever it is possible to bound the mixing probabilities, in our case $\frac{\pi_{11}}{\pi_{11}+\pi_{01}}, \frac{\pi_{01}}{\pi_{11}+\pi_{01}}$ and $\frac{\pi_{11}}{\pi_{11}+\pi_{10}}, \frac{\pi_{10}}{\pi_{11}+\pi_{10}}$, respectively, sharp bounds can be obtained for any parameter of the mixture components that respects stochastic dominance. We will use this fact to derive bounds for the average treatment effect (ATE) on various populations. A similar argument can be used to bound any parameter that respects stochastic dominance as, for instance, the quantile treatment effect (QTE).

3.1 Worst case bounds

Assume that the support of the potential outcomes is bounded, i.e., $Y(1), Y(0) \in [Y^{LB}, Y^{UB}]$, where Y^{LB}, Y^{UB} are the values at the lower and upper end of the support, respectively.This condition rules out infinite upper or lower bounds on the ATE in any population even without imposing restrictions other than Assumptions 1 and 2. In the absence of further assumptions Zhang and Rubin (2003) derive the worst case bounds of the ATE on the always selected (stratum 11), henceforth denoted as Δ_{11}, which are shown to be sharp in Imai (2008). In order to obtain their results, note that Table 3 provides us with the following equations:

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

such that

$$\pi_{01} \in [\max(0, P_{1|0} - P_{1|1}), \min(P_{1|0}, P_{0|1})]. \tag{2}$$

For the sake of brevity, let $\bar{Y}_{t,s} \equiv E(Y|T = t, S = s)$, i.e., the mean of Y given $T = t$ and $S = s$ (which is only observed for $S = 1$). Furthermore, let $F_{Y_{t,s}}(y) \equiv \Pr(Y \leq y|T = t, S = s)$ and $F^{-1}_{Y_{t,s}}(q) \equiv \inf\{y : F_{Y_{t,s}}(y) \geq q\}$, i.e., the conditional cdf and quantile function of Y given $T = t$ and $S = s$. Finally, let $\bar{Y}_{t,s}(\min|q) \equiv E(Y|T = t, S = s, y \leq F^{-1}_{Y_{t,s}}(q))$ and $\bar{Y}_{t,s}(\max|q) \equiv E(Y|T = t, S = s, y \geq F^{-1}_{Y_{t,s}}(1 - q))$. The upper and the lower bound of the ATE on the always selected, denoted as Δ^{UB}_{11} and Δ^{LB}_{11}, in Zhang and Rubin (2003), are

$$\begin{aligned} \Delta^{UB}_{11} &= \min_{\pi_{01}}[\bar{Y}_{1,1}(\max|(P_{1|0} - \pi_{01})/P_{1|1}) - \bar{Y}_{0,1}(\min|(P_{1|0} - \pi_{01})/P_{1|0})], \\ \Delta^{LB}_{11} &= \max_{\pi_{01}}[\bar{Y}_{1,1}(\min|(P_{1|0} - \pi_{01})/P_{1|1}) - \bar{Y}_{0,1}(\max|(P_{1|0} - \pi_{01})/P_{1|0})]. \end{aligned} \tag{3}$$

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 appendix, Δ_{11}^{UB} and Δ_{11}^{LB} can be simplified to

$$\begin{aligned}
\Delta_{11}^{UB} &= \bar{Y}_{1,1}(\max|(P_{1|0}-\pi_{01}^{\max})/P_{1|1}) - \bar{Y}_{0,1}(\min|(P_{1|0}-\pi_{01}^{\max})/P_{1|0}), \\
\Delta_{11}^{LB} &= \bar{Y}_{1,1}(\min|(P_{1|0}-\pi_{01}^{\max})/P_{1|1}) - \bar{Y}_{0,1}(\max|(P_{1|0}-\pi_{01}^{\max})/P_{1|0}),
\end{aligned} \tag{4}$$

where $\pi_{01}^{\max} \equiv \min(P_{1|0}, P_{0|1})$. The intuition of this result is that since we are able to bound any strata proportion, we can apply the findings in Horowitz and Manski (1995) to bound the mean of the always selected under treatment and non-treatment. The reason why one does not require numerical optimization is that the upper and lower bounds are maximized and minimized, respectively, when $\pi_{01} = \pi_{01}^{\max}$, see the appendix. Note that the bounds are only informative (i.e., tighter than $Y^{UB} - Y^{LB}$) if $P_{1|0} > P_{0|1}$, which has also been noticed by Lee (2009). This implies that $\pi_{11} > \pi_{00}$, i.e., that the share of always selected is larger than the share of never selected. If this is the case (and only then the bounds for the always selected are meaningful), then $(P_{1|0}-\pi_{01}^{\max})/P_{1|1} = (P_{1|0}-P_{0|1})/P_{1|1}$ and $(P_{1|0}-\pi_{01}^{\max})/P_{1|0} = (P_{1|0}-P_{0|1})/P_{1|0}$, such that the bound only depends on this ratio of observed proportions. Equivalent bounds can be derived keeping any other stratum proportion fixed, e.g. π_{11}, instead of π_{01}.

In contrast to previous work we will now also derive bounds for the compliers (stratum 10), the defiers (stratum 01), and the selected population. It is obvious from our previous discussion that the share of compliers in $o(1,1)$ is $\pi_{10}/(\pi_{11}+\pi_{10}) = (P_{1|1} - P_{1|0} + \pi_{01})/P_{1|1}$, i.e., the fraction of those who are not always selected. 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, denoted as Δ_{10}^{UB} and Δ_{10}^{LB}, are

$$\begin{aligned}\Delta_{10}^{UB} &= \bar{Y}_{1,1}(\max|(P_{1|1} - P_{1|0} + \pi_{01}^{\min})/P_{1|1}) - Y^{LB},\\ \Delta_{10}^{LB} &= \bar{Y}_{1,1}(\min|(P_{1|1} - P_{1|0} + \pi_{01}^{\min})/P_{1|1}) - Y^{UB}.\end{aligned} \tag{5}$$

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

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. Since 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, denoted as Δ_{01}^{UB} and Δ_{01}^{LB}, are

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

These bounds are only informative if $P_{1|0} - P_{1|1} > 0 \Rightarrow \pi_{01} > \pi_{10}$, i.e., 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 subsection.

Finally, we derive the worst case bounds for the selected population, which is a mixed population of always selected, compliers, and defiers. Their respective shares are given by

$$\begin{aligned}
\frac{2\cdot\pi_{11}}{2\pi_{11}+\pi_{10}+\pi_{01}} &= \frac{2\cdot(P_{1|0}-\pi_{01})}{P_{1|1}+P_{1|0}}, \\
\frac{\pi_{10}}{2\pi_{11}+\pi_{10}+\pi_{01}} &= \frac{P_{1|1}-P_{1|0}+\pi_{01}}{P_{1|1}+P_{1|0}}, \\
\frac{\pi_{01}}{2\pi_{11}+\pi_{10}+\pi_{01}} &= \frac{\pi_{01}}{P_{1|1}+P_{1|0}}.
\end{aligned}$$

Note that assuming the upper bound of the mean potential outcome under treatment for the always selected, $\bar{Y}_{1,1}(\max|P_{1|0}/P_{1|1})$ implies assuming the lower bound of the mean potential outcome under treatment for the compliers, $\bar{Y}_{1,1}(\min|1-P_{1|0}/P_{1|1})$, and vice versa, as the weighted average of both must always yield $\bar{Y}_{1,1}$. For the same reason, assuming the upper bound of the mean potential outcome under non-treatment for the always selected is equivalent to assuming the lower bound of the mean potential outcome under non-treatment for the defiers. In Appendix A.1.4 we use this fact to obtain unambiguous expressions for the bounds on the ATE on the selected population ($\Delta_{S=1}$) for some fixed defiers' share π_{01}. Furthermore, the appendix also shows that we need not optimize over the range of all the possible values of the defiers' share. In fact, setting $\pi_{01}=\pi_{01}^{\max}$ leads to sharp bounds on $\Delta_{S=1}$. Then, the upper bound is

$$\begin{aligned}
\Delta_{S=1}^{UB} &= \frac{2\cdot(P_{1|0}-\pi_{01}^{\max})}{P_{1|1}+P_{1|0}}\cdot(\bar{Y}_{1,1}(\max|(P_{1|0}-\pi_{01}^{\max})/P_{1|1})-\bar{Y}_{0,1}(\min|(P_{1|0}-\pi_{01}^{\max})/P_{1|0})) \\
&+ \frac{P_{1|1}-P_{1|0}+\pi_{01}^{\max}}{P_{1|1}+P_{1|0}}\cdot(\bar{Y}_{1,1}(\min|(P_{1|1}-P_{1|0}+\pi_{01}^{\max})/P_{1|1})-Y^{LB}) \\
&+ \frac{\pi_{01}^{\max}}{P_{1|1}+P_{1|0}}\cdot(Y^{UB}-\bar{Y}_{0,1}(\max|\pi_{01}^{\max}/P_{1|0}))
\end{aligned}$$

which, as simple calculus shows, can be expressed as

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

The lower bound is given by

$$\begin{aligned}
\Delta^{LB}_{S=1} &= \frac{2\cdot(P_{1|0}-\pi^{\max}_{01})}{P_{1|1}+P_{1|0}}\cdot(\bar{Y}_{1,1}(\min|(P_{1|0}-\pi^{\max}_{01})/P_{1|1})-\bar{Y}_{0,1}(\max|(P_{1|0}-\pi^{\max}_{01})/P_{1|0}))\\
&+ \frac{P_{1|1}-P_{1|0}+\pi^{\max}_{01}}{P_{1|1}+P_{1|0}}\cdot(\bar{Y}_{1,1}(\max|(P_{1|1}-P_{1|0}+\pi^{\max}_{01})/P_{1|1})-Y^{UB})\\
&+ \frac{\pi^{\max}_{01}}{P_{1|1}+P_{1|0}}\cdot(Y^{LB}-\bar{Y}_{0,1}(\min|\pi^{\max}_{01}/P_{1|0}))
\end{aligned}$$

which can be rewritten as

$$\begin{aligned}
\Delta^{LB}_{S=1} &= \frac{(P_{1|0}-\pi^{\max}_{01})}{P_{1|1}+P_{1|0}}\cdot(\bar{Y}_{1,1}(\min|(P_{1|0}-\pi^{\max}_{01})/P_{1|1})-\bar{Y}_{0,1}(\max|(P_{1|0}-\pi^{\max}_{01})/P_{1|0}))\\
&+ \frac{P_{1|1}}{P_{1|1}+P_{1|0}}\cdot\bar{Y}_{1,1}+\frac{\pi^{\max}_{01}}{P_{1|1}+P_{1|0}}\cdot Y^{LB}\\
&- \frac{P_{1|1}-P_{1|0}+\pi^{\max}_{01}}{P_{1|1}+P_{1|0}}\cdot Y^{UB}-\frac{P_{1|0}}{P_{1|1}+P_{1|0}}\cdot\bar{Y}_{0,1}.
\end{aligned}\tag{8}$$

Note that 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 selected 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 subsection. Another result worth noting is that the bounds on the selected population are always informative. E.g., if $P_{1|0} < P_{0|1}$, the bounds become

$$\begin{aligned}
\Delta^{UB}_{S=1} &= \frac{P_{1|1}}{P_{1|1}+P_{1|0}}\cdot\bar{Y}_{1,1}+\frac{P_{1|0}}{P_{1|1}+P_{1|0}}\cdot Y^{UB}-\frac{P_{1|1}}{P_{1|1}+P_{1|0}}\cdot Y^{LB}-\frac{P_{1|0}}{P_{1|1}+P_{1|0}}\cdot\bar{Y}_{0,1}\\
&= \frac{P_{1|1}}{P_{1|1}+P_{1|0}}\cdot(\bar{Y}_{1,1}-Y^{LB})+\frac{P_{1|0}}{P_{1|1}+P_{1|0}}\cdot(Y^{UB}-\bar{Y}_{0,1})
\end{aligned}$$

and

$$\begin{aligned}
\Delta^{LB}_{S=1} &= \frac{P_{1|1}}{P_{1|1}+P_{1|0}}\cdot\bar{Y}_{1,1}+\frac{P_{1|0}}{P_{1|1}+P_{1|0}}\cdot Y^{LB}-\frac{P_{1|1}}{P_{1|1}+P_{1|0}}\cdot Y^{UB}-\frac{P_{1|0}}{P_{1|1}+P_{1|0}}\cdot\bar{Y}_{0,1}\\
&= \frac{P_{1|1}}{P_{1|1}+P_{1|0}}\cdot(\bar{Y}_{1,1}-Y^{UB})+\frac{P_{1|0}}{P_{1|1}+P_{1|0}}\cdot(Y^{LB}-\bar{Y}_{0,1}),
\end{aligned}$$

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

3.2 Monotonicity

A commonly imposed assumption in the literature on partial identification of treatment effects under sample selection is weak monotonicity of selection w.r.t. the treatment:

Assumption 3:

$\Pr(S(1) \geq S(0)) = 1$ (monotonicity of selection).

In terms of the structural model in (1) this can be stated as

Assumption 3SM:

$\Pr(\varsigma(1, v) \geq \varsigma(0, v)) = 1 \; \forall v$ in the support of V.

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 $\Pr(S(0) \geq S(1)) = 1$ which implies that stratum 10 does not exist. As already mentioned before, assuming $\Pr(S(1) \geq S(0)) = 1$ (positive monotonicity) is only consistent with the data if $P_{1|1} - P_{1|0} \geq 0$ and $\Pr(S(0) \geq S(1)) = 1$ (negative monotonicity) if $P_{1|0} - P_{1|1} \geq 0$. 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. E.g., 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 such that defiers can be ruled out.

Monotonicity has been considered in Lee (2009), Zhang and Rubin (2003), and Zhang, Rubin, and Mealli (2008) to bound the ATE on the always selected (stratum 11) and in Lechner and Melly (2007) to derive bounds for the treated and selected population. Lee (2009) shows that the following bounds are sharp for the ATE on the always selected:

$$\begin{aligned} \Delta_{11}^{UB} &= \bar{Y}_{1,1}(\max|P_{1|0}/P_{1|1}) - \bar{Y}_{0,1}, \\ \Delta_{11}^{LB} &= \bar{Y}_{1,1}(\min|P_{1|0}/P_{1|1}) - \bar{Y}_{0,1}. \end{aligned} \tag{9}$$

The intuition of this result is that under monotonicity, $o(0,1)$ consists only of individuals belonging to stratum 11 such that $\bar{Y}_{0,1}$ is the mean potential outcome of the always selected under non-treatment. Furthermore, $P_{1|0} = \pi_{11}$. Therefore, the share of the always selected 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 selected 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 selected population and either the compliers or the defiers, without further assumptions. By introducing monotonicity we also identify meaningful bounds for the always selected, which turn out to be more informative than under the stochastic dominance assumption discussed further 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 it immediately follows that Δ_{10} is bounded by

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

It is obvious that 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 (2).

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

$$\begin{aligned}
\frac{2 \cdot \pi_{11}}{2 \cdot \pi_{11} + \pi_{10}} &= \frac{2 \cdot P_{1|0}}{P_{1|1} + P_{1|0}} \\
\frac{\pi_{10}}{2 \cdot \pi_{11} + \pi_{10}} &= \frac{P_{1|1} - P_{1|0}}{P_{1|1} + P_{1|0}}.
\end{aligned}$$

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

$$\begin{aligned}
\Delta_{S=1}^{UB} &= \frac{2 \cdot P_{1|0}}{P_{1|1} + P_{1|0}} \cdot (\bar{Y}_{1,1}(\max |P_{1|0}/P_{1|1}) - \bar{Y}_{0,1}) \\
&+ \frac{P_{1|1} - P_{1|0}}{P_{1|1} + P_{1|0}} \cdot (\bar{Y}_{1,1}(\min |1 - P_{1|0}/P_{1|1}) - Y^{LB}), \\
&= \frac{P_{1|0}}{P_{1|1} + P_{1|0}} \cdot \bar{Y}_{1,1}(\max |P_{1|0}/P_{1|1}) - 2 \cdot \frac{P_{1|0}}{P_{1|1} + P_{1|0}} \cdot \bar{Y}_{0,1} \\
&+ \frac{P_{1|1}}{P_{1|1} + P_{1|0}} \cdot \bar{Y}_{1,1} - \frac{P_{1|1} - P_{1|0}}{P_{1|1} + P_{1|0}} \cdot Y^{LB},
\end{aligned} \tag{11}$$

$$
\begin{aligned}
\Delta^{LB}_{S=1} &= \frac{2 \cdot P_{1|0}}{P_{1|1}+P_{1|0}} \cdot (\bar{Y}_{1,1}(\min|P_{1|0}/P_{1|1}) - \bar{Y}_{0,1}) \\
&+ \frac{P_{1|1}-P_{1|0}}{P_{1|1}+P_{1|0}} \cdot (\bar{Y}_{1,1}(\max|1-P_{1|0}/P_{1|1}) - Y^{UB}), \\
&= \frac{P_{1|0}}{P_{1|1}+P_{1|0}} \cdot \bar{Y}_{1,1}(\min|P_{1|0}/P_{1|1}) - 2 \cdot \frac{P_{1|0}}{P_{1|1}+P_{1|0}} \cdot \bar{Y}_{0,1} \\
&+ \frac{P_{1|1}}{P_{1|1}+P_{1|0}} \cdot \bar{Y}_{1,1} - \frac{P_{1|1}-P_{1|0}}{P_{1|1}+P_{1|0}} \cdot Y^{UB}. \qquad (12)
\end{aligned}
$$

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

3.3 Stochastic dominance

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

Assumption 4:

$\Pr(Y(t) \leq y|G = 11) \leq \Pr(Y(t) \leq y|G = 10), \quad t \in \{0, 1\},$

and

$\Pr(Y(t) \leq y|G = 11) \leq \Pr(Y(t) \leq y|G = 01), \quad t \in \{0, 1\}$ (stochastic dominance).

I.e., the potential outcome among the always selected 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.[2] Taking the evaluation of the returns to a job training as example, it implies that the always selected have potential wages that are at least as high as the ones of other groups.

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

To justify Assumption 4, note that the always selected are employed irrespective of the training. Therefore, they are likely to be more motivated and/or able than other populations. Zhang, Rubin, and Mealli (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 selected 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 selected:

$$\begin{aligned}\Delta_{11}^{UB} &= \bar{Y}_{1,1}(\max|(P_{1|0}-\pi_{01}^{\max})/P_{1|1}) - \bar{Y}_{0,1},\\ \Delta_{11}^{LB} &= \bar{Y}_{1,1} - \bar{Y}_{0,1}(\max|(P_{1|0}-\pi_{01}^{\max})/P_{1|0}).\end{aligned} \tag{13}$$

As $E[Y(t)|G=11] \geq E[Y(t)|G=10]$, $E[Y(t)|G=11] \geq E[Y(t)|G=01]$ 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 $E[Y(0)|G=11]$, respectively. Thus, Assumption 4 is likely to shrink the worst case bounds because $\bar{Y}_{0,1} \geq \bar{Y}_{0,1}(\min|(P_{1|0}-\pi_{01}^{\max})/P_{1|0})$ and $\bar{Y}_{1,1} \geq \bar{Y}_{1,1}(\min|(P_{1|0}-\pi_{01}^{\max})/P_{1|1})$. Note that width of the bounds is maximized if the share of the always selected is smaller than the one of the never selected. Then, $\bar{Y}_{1,1}(\max|(P_{1|0} - \pi_{01}^{\max})/P_{1|1}) = \bar{Y}_{1,1}(\max|0)$ and $\bar{Y}_{0,1}(\max|(P_{1|0}-\pi_{01}^{\max})/P_{1|0}) = \bar{Y}_{0,1}(\max|0)$ such 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:

$$\begin{aligned}\Delta_{10}^{UB} &= \bar{Y}_{1,1} - Y^{LB},\\ \Delta_{10}^{LB} &= \min_{\pi_{01}}\left[\bar{Y}_{1,1}(\min|(P_{1|1}-P_{1|0}+\pi_{01})/P_{1|1}) - \bar{Y}_{0,1}(\max|(P_{1|0}-\pi_{01})/P_{1|0})\right].\end{aligned} \tag{14}$$

The intuition is that any mean potential outcome of the compliers is at best as high as that of the always selected, such 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 to be tighter than the worst case bounds since $\bar{Y}_{1,1} \leq \bar{Y}_{1,1}(\max|(P_{1|1} - P_{1|0} + \pi_{01})/P_{1|1})$ and $\bar{Y}_{0,1}(\max|(P_{1|0} - \pi_{01})/P_{1|0}) \leq Y^{UB}$. In particular, stochastic dominance bears considerable identifying power for the lower bound of the effect, since it does not depend on Y^{UB} anymore. This is a relevant result 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 since $\bar{Y}_{1,1}(\min|(P_{1|1} - P_{1|0} + \pi_{01})/P_{1|1})$ 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

$$\begin{aligned}
\Delta_{01}^{UB} &= \max_{\pi_{01}} \left[\bar{Y}_{1,1}(\max|(P_{1|0} - \pi_{01})/P_{1|1}) - \bar{Y}_{0,1}(\min|\pi_{01}/P_{1|0})\right], \\
\Delta_{01}^{LB} &= Y^{LB} - \bar{Y}_{0,1}. \qquad (15)
\end{aligned}$$

As for the compliers, any mean potential outcome of the defiers can be at best as high as the one of the always selected such 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 since $\bar{Y}_{1,1}(\max|(P_{1|0} - \pi_{01})/P_{1|1}) \leq Y^{UB}$ and $\bar{Y}_{0,1} \leq \bar{Y}_{0,1}(\max|\pi_{01}/P_{1|0})$. Since $\bar{Y}_{1,1}(\max|(P_{1|0} - \pi_{01})/P_{1|1})$ 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 selected population are identified by

$$\begin{aligned}
\Delta_{S=1}^{UB} &= \frac{P_{1|0}}{P_{1|1} + P_{1|0}} \cdot \bar{Y}_{1,1}(\max|(P_{1|0} - \pi_{01}^{\max})/P_{1|1}) - \frac{2 \cdot P_{1|0} - \pi_{01}^{\max}}{P_{1|1} + P_{1|0}} \cdot \bar{Y}_{0,1} \\
&+ \frac{P_{1|1}}{P_{1|1} + P_{1|0}} \cdot \bar{Y}_{1,1} - \frac{P_{1|1} - P_{1|0} + \pi_{01}^{\max}}{P_{1|1} + P_{1|0}} \cdot Y^{LB}, \\
\Delta_{S=1}^{LB} &= \frac{P_{1|1} + P_{1|0} - \pi_{01}^{\max}}{P_{1|1} + P_{1|0}} \cdot \bar{Y}_{1,1} - \frac{P_{1|0}}{P_{1|1} + P_{1|0}} \cdot \bar{Y}_{0,1}(\max|(P_{1|0} - \pi_{01}^{\max})/P_{1|0})) \\
&- \frac{P_{1|0}}{P_{1|1} + P_{1|0}} \cdot \bar{Y}_{0,1} + \frac{\pi_{01}^{\max}}{P_{1|1} + P_{1|0}} \cdot Y^{LB}. \qquad (16)
\end{aligned}$$

For both the upper and the lower bound of $\Delta_{S=1}$, stochastic dominance eliminates either Y^{UB} or Y^{LB} present in the worst case scenario. Intuitively, the identification region must shrink since the bounds for always selected, compliers, and defiers become narrower. E.g., in the case that the never selected outnumber the always selected the bounds become

$$\begin{aligned}
\Delta_{S=1}^{UB} &= \frac{P_{1|1}}{P_{1|1}+P_{1|0}} \cdot (\bar{Y}_{1,1} - Y^{LB}) + \frac{P_{1|0}}{P_{1|1}+P_{1|0}} \cdot (Y^{UB} - \bar{Y}_{0,1}), \\
\Delta_{S=1}^{LB} &= \frac{P_{1|1}}{P_{1|1}+P_{1|0}} \cdot \bar{Y}_{1,1} + \frac{P_{1|0}}{P_{1|1}+P_{1|0}} \cdot (Y^{LB} - \bar{Y}_{0,1} - Y^{UB}). \qquad (17)
\end{aligned}$$

Interestingly, these bounds are are tighter than the ones on Δ_{11}. Once again, we obtain more informative bounds for the selected population than for the always selected if $\pi_{00} > \pi_{11}$.

3.4 Monotonicity and stochastic dominance

In the subsequent discussion, we investigate the identifying power of a combination of Assumptions 3 and 4. This was first considered by Zhang and Rubin (2003) who derive the following bounds for the ATE on the always selected which are shown to be sharp by Imai (2008):

$$\begin{aligned}
\Delta_{11}^{UB} &= \bar{Y}_{1,1}(\max |P_{1|0}/P_{1|1}) - \bar{Y}_{0,1}, \\
\Delta_{11}^{LB} &= \bar{Y}_{1,1} - \bar{Y}_{0,1}. \qquad (18)
\end{aligned}$$

These bounds are a simplification of those under stochastic dominance for the special case that $\pi_{01} = 0$. The upper bound is the same as under monotonicity and is, thus, not affected by additionally assuming stochastic dominance. The latter has no further impact on 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 selected under treatment.

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

$$
\begin{aligned}
\Delta_{10}^{UB} &= \bar{Y}_{1,1} - Y^{LB}, \\
\Delta_{10}^{LB} &= \bar{Y}_{1,1}(\min |1 - P_{1|0}/P_{1|1}) - \bar{Y}_{0,1}.
\end{aligned} \tag{19}
$$

The upper bound is the same as under stochastic dominance and is not affected by adding monotonicity. The reason is that 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 $\pi_{01} \geq 0$, $\pi_{01} = 0$ 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 selected population are identified by

$$
\begin{aligned}
\Delta_{S=1}^{UB} &= \frac{P_{1|0}}{P_{1|1} + P_{1|0}} \cdot \bar{Y}_{1,1}(\max |P_{1|0}/P_{1|1}) - 2 \cdot \frac{P_{1|0}}{P_{1|1} + P_{1|0}} \cdot \bar{Y}_{0,1} \\
&+ \frac{P_{1|1}}{P_{1|1} + P_{1|0}} \cdot \bar{Y}_{1,1} - \frac{P_{1|1} - P_{1|0}}{P_{1|1} + P_{1|0}} \cdot Y^{LB}, \\
\Delta_{S=1}^{LB} &= \frac{2P_{1|0}}{P_{1|1} + P_{1|0}} \cdot (\bar{Y}_{1,1} - \bar{Y}_{0,1}) + \frac{P_{1|1} - P_{1|0}}{P_{1|1} + P_{1|0}} \cdot (\bar{Y}_{1,1} - \bar{Y}_{0,1}) \\
&= \bar{Y}_{1,1} - \bar{Y}_{0,1}.
\end{aligned} \tag{20}
$$

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 selected 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 selected 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 selected.

3.5 Monotone treatment response

As the last identifying assumption considered, we briefly discuss the so called "monotone treatment response" assumption suggested by Manski (1997):

Assumption 5:
$\Pr(Y(1) \geq Y(0)) = 1$ (montone treatment response).

Assumption 5 restricts the effect of the treatment to be non-negative implying that the lower bound under any other assumptions discussed so far may not be smaller than zero. Apart from Angrist, Bettinger, and Kremer (2006), who use this assumption together with Assumption 3 to bound the ATE of private school vouchers, monotone treatment response has received little attention in the literature on partial identification under sample selection.[3] This may be due to the limited attractiveness of a priori restricting the direction of the effect when the latter is unknown to the researcher. We will, therefore, not invoke Assumption 5 in the application discussed further below.

4 Estimation

While the focus of our paper lies on identification, this section briefly sketches estimation, which is based on the sample analogs of the bounds derived under the various assumptions. To this

[3]However, it has been very influential elsewhere, see for instance Gerfin and Schellhorn (2006), González (2005), and Kreider and Hill (2009) for empirical applications in health and labor economics. Furthermore, Blanco, Flores, and Flores-Lagunes (2011) consider a somewhat related assumption which restricts the ATE on the always selected to be non-negative.

end, we define the following sample parameters:

$$
\begin{aligned}
\hat{P}_{1|1} &\equiv \frac{\sum_{i=1}^{n} S_i \cdot T_i}{\sum_{i=1}^{n} T_i}, \qquad \hat{P}_{0|1} \equiv 1 - \frac{\sum_{i=1}^{n} S_i \cdot T_i}{\sum_{i=1}^{n} T_i}, \\
\hat{P}_{1|0} &\equiv \frac{\sum_{i=1}^{n} S_i \cdot (1-T_i)}{\sum_{i=1}^{n} (1-T_i)}, \qquad \hat{P}_{0|0} \equiv 1 - \frac{\sum_{i=1}^{n} S_i \cdot (1-T_i)}{\sum_{i=1}^{n} (1-T_i)}, \\
\hat{\bar{Y}}_{1,1} &\equiv \frac{\sum_{i=1}^{n} Y_i \cdot S_i \cdot T_i}{\sum_{i=1}^{n} S_i \cdot T_i}, \qquad \hat{\bar{Y}}_{0,1} \equiv \frac{\sum_{i=1}^{n} Y_i \cdot S_i \cdot (1-T_i)}{\sum_{i=1}^{n} S_i \cdot (1-T_i)}, \\
\hat{\bar{Y}}_{t,s}(\max|q) &\equiv \frac{\sum_{i=1}^{n} Y_i \cdot I\{S_i = s\} \cdot I\{T_i = t\} \cdot I\{Y \geq \hat{y}_{1-q}\}}{\sum_{i=1}^{n} I\{S_i = s\} \cdot I\{T_i = t\} \cdot I\{Y \geq \hat{y}_{1-q}\}}, \\
\hat{\bar{Y}}_{t,s}(\min|q) &\equiv \frac{\sum_{i=1}^{n} Y_i \cdot I\{S_i = s\} \cdot I\{T_i = t\} \cdot I\{Y \leq \hat{y}_{q}\}}{\sum_{i=1}^{n} I\{S_i = s\} \cdot I\{T_i = t\} \cdot I\{Y \leq \hat{y}_{q}\}}, \\
\hat{y}_q &\equiv \min\left\{ y : \frac{\sum_{i=1}^{n} S_i \cdot T_i \cdot I\{Y_i \leq y\}}{\sum_{i=1}^{n} S_i \cdot T_i} \geq q \right\},
\end{aligned}
$$

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. E.g., 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 selected population under both monotonicity and stochastic dominance directly follows from the results of Lee (2009). To see this, first consider the estimators of $\Delta_{11}^{UB}, \Delta_{11}^{LB}$ under monotonicity alone:

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

In the 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). As argued therein, it is obvious that it suffices to show the desirable properties for $\hat{\bar{Y}}_{1,1}(\max|\hat{P}_{1|0}/\hat{P}_{1|1})$ and

$\hat{\hat{Y}}_{1,1}(\max|\hat{P}_{1|0}/\hat{P}_{1|1})$ (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 $\hat{\hat{Y}}_{0,1}$.

Now consider the estimators for the compliers under monotonicity:

$$\begin{aligned}\hat{\Delta}_{10}^{UB} &= \hat{\hat{Y}}_{1,1}(\max|1-P_{1|0}/P_{1|1}) - Y^{LB}, \\ \hat{\Delta}_{10}^{LB} &= \hat{\hat{Y}}_{1,1}(\min|1-P_{1|0}/P_{1|1}) - Y^{UB}. \qquad (21)\end{aligned}$$

Y^{LB}, Y^{UB} are constants not relevant for the properties of the estimators. Furthermore, note that the problem of estimating $\hat{\hat{Y}}_{1,1}(\max|1-P_{1|0}/P_{1|1})$ is symmetric to $\hat{\hat{Y}}_{1,1}(\max|\hat{P}_{1|0}/\hat{P}_{1|1})$ (and $\hat{\hat{Y}}_{1,1}(\min|1-P_{1|0}/P_{1|1})$ to $\hat{\hat{Y}}_{1,1}(\max|\hat{P}_{1|0}/\hat{P}_{1|1}))$. 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 $\hat{\Delta}_{S=1}^{UB}, \hat{\Delta}_{S=1}^{LB}$, as the selected population is just a weighted average of the always selected 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.

5 Application

In this section, we use our methods to re-evaluate the school voucher experiment of Angrist, Bettinger, and Kremer (2006). As mentioned before, the authors investigate the effects of school vouchers provided to high school students in the course of Colombia's 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 such that Assumption 2 appears likely to hold. The experimental estimates in Angrist, Bettinger, and Kremer (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 1223 students) of the 4044 applicants actually took the test. Therefore, the experimental estimates might be flawed by selection bias. E.g., if the treatment positively affects the likelihood to take the test such 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, Bettinger, and Kremer (2006) use both censored regression to control for sample selection and derive nonparametric bounds on the ATE of the always selected 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 selected, compliers, and the selected 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 selected 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 last section. 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%:

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

where $\hat{\Delta}^{LB}, \hat{\Delta}^{UB}$ are the estimated bounds and $\hat{\sigma}^{LB}, \hat{\sigma}^{UB}$ denote their respective estimated standard errors. We compute the latter by bootstrapping the original sample 1999 times and estimating $\hat{\Delta}^{LB}, \hat{\Delta}^{UB}$ 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 selected is larger than the upper bound on the share of any other population and in particular than the one of the always selected. 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 such 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 4: Estimated (bounds on the) proportions of latent strata

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

Table 5 presents the results for the always selected, compliers, and the selected 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 selected and very wide for any other population. Invoking monotonicity narrows the bounds substantially for the always selected and the selected population, even though the identification region still includes a zero effect. As discussed before, monotonicity has no identifying power for the compliers as a zero proportion of π_{01} implies the widest bounds possible.

The stochastic dominance assumption entails narrower bounds than the worst case scenario for all three populations. However, for the always selected, 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 selected and the selected 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 selected 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 3.

Table 5: ATE estimates and confidence intervals

Assumptions	Always selected	Compliers	Selected
Worst case bounds	[-31.000, 32.000]	[-24.083, 24.799]	[-16.632, 17.368]
	Not informative	(-25.658, 26.106)	(-16.891, 17.628)
Monotonicity	[-0.942, 2.218]	[-24.083, 24.799]	[-1.805, 3.115]
	(-1.731, 2.978)	(-25.658, 26.106)	(-2.824, 4.122)
Stochastic dominance	[-13.396, 17.079]	[-13.439, 17.604]	[-8.706, 17.368]
	(-14.666, 18.755)	(-14.566, 20.852)	(-9.403, 18.068)
Mon. + stoch. dom.	[0.683, 2.218]	[-7.004, 17.604]	[0.683, 3.115]
	(0.152, 2.978)	(-8.626, 17.932)	(0.152, 4.122)

Note: Bounds in square brackets and confidence intervals in round brackets.
Confidence intervals are based on 1999 bootstraps.

All in all our results give support to the conclusions of Angrist, Bettinger, and Kremer (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.

6 Conclusion

This paper discusses the partial identification of average treatment effects (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 selected", whose outcomes are observed irrespective of the treatment received. Here, we also derived sharp bounds for other populations such as the "compliers" (selected under treatment, not selected under non-treatment) and the selected 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 selected population, such 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 selected over those of others) assumptions may bear considerable identifying power even for populations whose outcomes are, in contrast to the always selected, only observed in one treatment state. In particular, it has been shown that the lower bound of the ATE on the selected population coincides with the lower bound for the always selected. 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 paper 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 behavior 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 selected is smaller than the one of never selected, bounds on the always selected are not informative if we do not assume monotonicity of selection in the treatment. In contrast, we can still bound the ATE on the selected population. This might be hard to see from the equations characterizing a structural model.

A Appendix

A.1 Worst case scenario

A.1.1 Proof of the sharpness of the simplified bounds on Δ_{11}

The bounds in Zhang and Rubin (2003) are

$$\Delta_{11}^{UB} = \min_{\pi_{01}}[\bar{Y}_{1,1}(\max|(P_{1|0}-\pi_{01})/P_{1|1})-\bar{Y}_{0,1}(\min|(P_{1|0}-\pi_{01})/P_{1|0})], \quad (A.1)$$

$$\Delta_{11}^{LB} = \max_{\pi_{01}}[\bar{Y}_{1,1}(\min|(P_{1|0}-\pi_{01})/P_{1|1})-\bar{Y}_{0,1}(\max|(P_{1|0}-\pi_{01})/P_{1|0})]. \quad (A.2)$$

Imai (2008) shows that these bounds are sharp, but they can be simplified such that optimization is not required: $\bar{Y}_{1,1}(\max|(P_{1|0}-\pi_{01})/P_{1|1}) \equiv E(Y|T=1,S=1,y \geq F^{-1}_{Y_{1,1}}(1-(P_{1|0}-\pi_{01})/P_{1|1}))$ is maximized when $1-\frac{P_{1|0}-\pi_{01}}{P_{1|1}} = \frac{P_{1|1}-P_{1|0}+\pi_{01}}{P_{1|1}}$, is maximized. Since $\frac{P_{1|1}-P_{1|0}+\pi_{01}}{P_{1|1}}$ is increasing in π_{01}, it is maximized when $\pi_{01}=\pi_{01}^{\max}$. Similarly $\bar{Y}_{0,1}(\min|(P_{1|0}-\pi_{01})/P_{1|0})$ is minimized when $\frac{P_{1|0}-\pi_{01}}{P_{1|0}}$, which is decreasing in π_{01}, is minimized, namely when $\pi_{01}=\pi_{01}^{\max}$. Noticing that $\frac{P_{1|0}-\pi_{01}}{P_{1|1}}$ and $1-\frac{P_{1|0}-\pi_{01}}{P_{1|0}} = \frac{\pi_{01}}{P_{1|0}}$ are decreasing and increasing functions of π_{01}, respectively, ends the proof.

A.1.2 Proof of the sharpness of the bounds on Δ_{10}

Lemma 1 together with Proposition 1 in Imai (2008) shows that $\bar{Y}_{1,1}(\max|(P_{1|1}-P_{1|0}+\pi_{01})/P_{1|1})$ and $\bar{Y}_{1,1}(\min|(P_{1|1}-P_{1|0}+\pi_{01})/P_{1|1})$ are the sharp upper and lower bounds of $E(Y|T=1,G=10)$. Since the sampling process does not impose any restrictions on the distribution of Y given $T=0$ and $G=10$ for a fixed value of π_{01}, the bounds are sharp. Finally, since π_{01} is unknown, the bounds are obtained by maximizing (minimizing) the upper (the lower) bound w.r.t. to its admissible values. However, as $1-\frac{P_{1|1}-P_{1|0}+\pi_{01}}{P_{1|1}}$ and $\frac{P_{1|1}-P_{1|0}+\pi_{01}}{P_{1|1}}$ are decreasing and increasing in π_{01},respectively, $\pi_{01}^{\min}$ is the optimal choice in both cases.

A.1.3 Proof of the sharpness of the bounds on Δ_{01}

Lemma 1 together with Proposition 1 in Imai (2008) shows that $\bar{Y}_{0,1}(\max|\pi_{01}/P_{1|0})$ and $\bar{Y}_{0,1}(\min|\pi_{01}/P_{1|0})$ are the sharp upper and lower bounds of $E(Y|T=0,G=01)$. Since the sampling process does not impose any restriction on the distribution of Y given $T=1$ and $G=01$ for a fixed value of π_{01}, the bounds are sharp. Finally, since π_{01} is unknown, the bounds are obtained by maximizing (minimizing) the upper (the lower) bound w.r.t. to its admissible values. However, since $1-\frac{\pi_{01}}{P_{1|0}}$ and $\frac{\pi_{01}}{P_{1|0}}$ are decreasing and increasing in π_{01}, respectively, $\pi_{01}^{\min}$ is the optimal choice in both cases.

A.1.4 Proof of the sharpness of the bounds on $\Delta_{S=1}$

We will show the sharpness of the upper bound, the proof for the lower bound is analogous. First of all, notice that if w is a random variable which is distributed as a two components mixture

$$f(w) = p \cdot f(w_1) + (1-p) \cdot f(w_2) \;\; p \in [0,1],$$

then

$$E(w) = p \cdot E(w|w \geq F_w^{-1}(1-p)) + (1-p) \cdot E(w|w \leq F_w^{-1}(1-p)), \tag{A.3}$$

where $E(w|w \geq F_w^{-1}(1-p))$ is the upper bound of $E(w_1)$ and $E(w|w \leq F_w^{-1}(1-p))$ is the lower bound of $E(w_2)$. Given (A.3), the observed outcomes can be written as

$$\bar{Y}_{1,1} = \frac{P_{1|0} - \pi_{01}}{P_{1|1}} \cdot \bar{Y}_{1,1}(\max|(P_{1|0} - \pi_{01})/P_{1|1}) + \frac{P_{1|1} - P_{1|0} + \pi_{01}}{P_{1|1}} \cdot \bar{Y}_{1,1}(\min|(P_{1|1} - P_{1|0} + \pi_{01})/P_{1|1}) \tag{A.4}$$

or

$$\bar{Y}_{1,1} = \frac{P_{1|0} - \pi_{01}}{P_{1|1}} \cdot \bar{Y}_{1,1}(\min|(P_{1|0} - \pi_{01})/P_{1|1}) + \frac{P_{1|1} - P_{1|0} + \pi_{01}}{P_{1|1}} \cdot \bar{Y}_{1,1}(\max|(P_{1|1} - P_{1|0} + \pi_{01})/P_{1|1}) \tag{A.5}$$

and

$$\bar{Y}_{0,1} = \frac{P_{1|0} - \pi_{01}}{P_{1|0}} \cdot \bar{Y}_{0,1}(\min|(P_{1|0} - \pi_{01})/P_{1|0}) + \frac{\pi_{01}}{P_{1|0}} \cdot \bar{Y}_{0,1}(\max|\pi_{01}/P_{1|0}) \tag{A.6}$$

or

$$\bar{Y}_{0,1} = \frac{P_{1|0} - \pi_{01}}{P_{1|0}} \cdot \bar{Y}_{0,1}(\max|(P_{1|0} - \pi_{01})/P_{1|0}) + \frac{\pi_{01}}{P_{1|0}} \cdot \bar{Y}_{0,1}(\min|\pi_{01}/P_{1|0}). \tag{A.7}$$

Moreover, notice that

$$\Delta_{S=1} = \frac{2 \cdot (P_{1|0} - \pi_{01})}{P_{1|1} + P_{1|0}} \cdot \Delta_{11} + \frac{P_{1|1} - P_{1|0} + \pi_{01}}{P_{1|1} + P_{1|0}} \cdot \Delta_{10} + \frac{\pi_{01}}{P_{1|1} + P_{1|0}} \cdot \Delta_{01} \tag{A.8}$$

For the upper bound, substituting Δ_{11} by Δ_{11}^{UB}, Δ_{10} by Δ_{10}^{UB} and Δ_{01} by Δ_{01}^{UB} in (A.8) would give a sharp upper bound on $\Delta_{S=1}$. However, such a bound would contradict (A.3) since it is impossible to have the upper bounds for the always selected and the compliers and the lower bounds for the always selected and the defiers at the same time in the mixture. This, however, shows that the admissible sharp upper bound would be the maximum of

$$\begin{aligned}
\Delta^{UB}_{S=1} &= \frac{2\cdot(P_{1|0}-\pi_{01})}{P_{1|1}+P_{1|0}}\cdot(\bar{Y}_{1,1}(\text{min or max}\,|(P_{1|0}-\pi_{01})/P_{1|1})-\bar{Y}_{0,1}(\text{min or max}\,|(P_{1|0}-\pi_{01})/P_{1|0}))\\
&+ \frac{P_{1|1}-P_{1|0}+\pi_{01}}{P_{1|1}+P_{1|0}}\cdot(\bar{Y}_{1,1}(\text{min or max}\,|(P_{1|1}-P_{1|0}+\pi_{01})/P_{1|1})-Y^{LB})\\
&+ \frac{\pi_{01}}{P_{1|1}+P_{1|0}}\cdot(Y^{UB}-\bar{Y}_{0,1}(\text{min or max}\,|\pi_{01}/P_{1|0})).
\end{aligned}\tag{A.9}$$

From (A.4),(A.5), (A.6) and (A.7) we have, respectively,

$$\begin{aligned}
\bar{Y}_{1,1}(\min|(P_{1|1}-P_{1|0}+\pi_{01})/P_{1|1}) &= \frac{P_{1|1}}{P_{1|1}-P_{1|0}+\pi_{01}}\cdot\bar{Y}_{1,1}-\frac{P_{1|0}-\pi_{01}}{P_{1|1}-P_{1|0}+\pi_{01}}\\
&\cdot\ \bar{Y}_{1,1}(\max|(P_{1|0}-\pi_{01})/P_{1|1}) \qquad \text{(A.10)}\\
\bar{Y}_{1,1}(\max|(P_{1|1}-P_{1|0}+\pi_{01})/P_{1|1}) &= \frac{P_{1|1}}{P_{1|1}-P_{1|0}+\pi_{01}}\cdot\bar{Y}_{1,1}-\frac{P_{1|0}-\pi_{01}}{P_{1|1}-P_{1|0}+\pi_{01}}\\
&\cdot\ \bar{Y}_{1,1}(\min|(P_{1|0}-\pi_{01})/P_{1|1}) \qquad \text{(A.11)}\\
\bar{Y}_{0,1}(\min|\pi_{01}/P_{1|0}) &= \frac{P_{1|0}}{\pi_{01}}\cdot\bar{Y}_{0,1}-\frac{P_{1|0}-\pi_{01}}{\pi_{01}}\\
&\cdot\ \bar{Y}_{0,1}(\max|(P_{1|0}-\pi_{01})/P_{1|0}) \qquad \text{(A.12)}\\
\bar{Y}_{0,1}(\max|\pi_{01}/P_{1|0}) &= \frac{P_{1|0}}{\pi_{01}}\cdot\bar{Y}_{0,1}-\frac{P_{1|0}-\pi_{01}}{\pi_{01}}\\
&\cdot\ \bar{Y}_{0,1}(\min|(P_{1|0}-\pi_{01})/P_{1|0}) \qquad \text{(A.13)}
\end{aligned}$$

Substituting these expressions in (A.9), we obtain after some simple algebra

$$\begin{aligned}
\Delta^{UB}_{S=1} &= \frac{P_{1|0}-\pi_{01}}{P_{1|1}+P_{1|0}}\cdot(\bar{Y}_{1,1}(\text{min or max}\,|(P_{1|0}-\pi_{01})/P_{1|1})-\bar{Y}_{0,1}(\text{min or max}\,|(P_{1|0}-\pi_{01})/P_{1|0}))\\
&+ \frac{P_{1|1}}{P_{1|1}+P_{1|0}}\cdot\bar{Y}_{1,1}-\frac{P_{1|1}-P_{1|0}+\pi_{01}}{P_{1|1}+P_{1|0}}\cdot Y^{LB}+\frac{\pi_{01}}{P_{1|1}+P_{1|0}}\cdot Y^{UB}-\frac{P_{1|0}}{P_{1|1}+P_{1|0}}\cdot\bar{Y}_{0,1}.
\end{aligned}$$

This is maximized for $\bar{Y}_{1,1}(\max|(P_{1|0}-\pi_{01})/P_{1|1})-\bar{Y}_{0,1}(\min|(P_{1|0}-\pi_{01})/P_{1|0})$. For a given value of π_{01} this bound is sharp. Finally, we again need to maximize w.r.t. π_{01}. However, we will show that $\Delta^{UB}_{S=1}$ is maximized for $\pi_{01}=\pi_{01}^{\max}$. Indeed, by taking its derivative w.r.t. π_{01}, defining $\bar{Y}_{1,1}(\max|(P_{1|0}-\pi_{01})/P_{1|1})-\bar{Y}_{0,1}(\min|(P_{1|0}-\pi_{01})/P_{1|0})\equiv\Delta Y$, and performing some simple algebra one obtains

$$Y^{UB}-Y^{LB}+P_{1|0}\cdot\frac{\partial\Delta Y}{\partial\pi_{01}}\geq\pi_{01}\cdot\frac{\partial\Delta Y}{\partial\pi_{01}}+\Delta Y. \tag{A.14}$$

which is always satisfied, because $Y^{UB}-Y^{LB}$ is generally larger than ΔY and $P_{1|0}\geq\pi_{01}$. (A.14) holds as an equality only if $\pi_{01}=P_{1|0}=\pi_{01}^{\max}$. This ends the proof.

A.2 Monotonicity

Any bounds derived under monotonicity are special cases of the worst case bounds given $\pi_{01} = 0$ and, therefore, they are sharp.

A.3 Stochastic dominance

First of all, we will prove the following lemma:

Lemma 1: *If w is a random variable which is distributed as a two components mixture*

$$f(w) = p \cdot f(w_1) + (1-p) \cdot f(w_2) \quad p \in [\lambda_0, \lambda_1],$$

and $F(w_1 \leq t) \geq F(w_2 \leq t)$ $\forall t \in (-\infty, +\infty)$, and if $g(\cdot)$ is a function that respect stochastic dominance, the following bounds are sharp:

$$g(F(w_1 \leq t)) \in [g(L_{\lambda_1}), g(U)],$$

where

$$\begin{aligned} L_\gamma &= \begin{cases} \frac{F(w \leq t)}{\gamma} & \text{if } t < F_w^{-1}(\gamma), \\ 1 & \text{if } t \geq F_w^{-1}(\gamma), \end{cases} \\ U &= F(w \leq t). \end{aligned}$$

Proof: The lower bound is sharp because of Lemma 1 in Imai (2008). Similar to Lemma 2 in Imai (2008), the identification region of $F(w_1 \leq t)$ is $\Psi_{w_1} \equiv \left\{ \frac{F(w \leq t) - (1-\lambda_1)\psi_{w_2}}{\lambda_1} : \psi_{w_2} \leq F(w \leq t) \right\}$ such that $U \equiv F(w \leq t) \leq \psi_{w_1}$ stochastically dominates any admissible value of the identification region of $F(w_1 \leq t)$. This proves the sharpness of the upper bound.

A.3.1 Proof of the sharpness of the bounds on Δ_{10}

Lemma 1 shows that $\bar{Y}_{1,1}$ and $\bar{Y}_{1,1}(\min |(P_{1|1} - P_{1|0} + \pi_{01})/P_{1|1})$ are the sharp upper and lower bounds on $E(Y|T = 1, G = 10)$. Under Assumption 4 the lower bound of $E(Y|T = 0, G = 10)$ remains Y^{LB}. On the other hand, the upper bound cannot be larger than $\bar{Y}_{0,1}(\max |(P_{1|0} - \pi_{01})/P_{1|0})$. This implies that the upper bound is sharp while the lower bound is sharp for any fixed value of π_{01}. The sharp lower bound is obtained minimizing over all possible values of π_{01}.

A.3.2 Proof of the sharpness of the bounds on Δ_{01}

Lemma 1 shows that $\bar{Y}_{0,1}$ and $\bar{Y}_{0,1}(\min|\pi_{01}^{\min}/P_{1|0})$ are the sharp upper and lower bounds for $E(Y|T=0,G=01)$. Under Assumption 4 the lower bound of $E(Y|T=1,G=01)$ remains Y^{LB}. On the other hand, the upper bound cannot be larger than $\bar{Y}_{1,1}(\min|(P_{1|0}-\pi_{01})/P_{1|1})$. This implies that the lower bound is sharp while the upper bound is sharp for any fixed value of π_{01}. The sharp upper bound is obtained maximizing over all possible values of π_{01}.

A.3.3 Proof of the sharpness of the bounds on $\Delta_{S=1}$

We only prove the sharpness of the upper bound, the proof for the lower bound is analogous. The sharp upper bound would be obtained substituting Δ_{11} with Δ_{11}^{UB}, Δ_{10} with Δ_{10}^{UB} and Δ_{01} with Δ_{01}^{UB} in (16). Again, such a bound would contradict (A.3).

Therefore, we have four admissible solutions:

1. Δ_{11}^{UB}, $\bar{Y}_{1,1}(\min|(P_{1|1}-P_{1|0}+\pi_{01})/P_{1|1})-Y^{LB}$ and $\bar{Y}_{1,1}-\bar{Y}_{0,1}$. After some algebra we have

$$\begin{aligned}\Delta_{S=1}^{UB1} &= \frac{P_{1|0}}{P_{1|1}+P_{1|0}}\cdot\bar{Y}_{1,1}(\max|(P_{1|0}-\pi_{01})/P_{1|1})-\frac{2\cdot P_{1|0}-\pi_{01}}{P_{1|1}+P_{1|0}}\cdot\bar{Y}_{0,1}\\ &+ \frac{P_{1|1}}{P_{1|1}+P_{1|0}}\cdot\bar{Y}_{1,1}-\frac{P_{1|1}-P_{1|0}+\pi_{01}}{P_{1|1}+P_{1|0}}\cdot Y^{LB}\end{aligned}$$

2. $\bar{Y}_{1,1}-\bar{Y}_{0,1}$, Δ_{10}^{UB} and $\bar{Y}_{1,1}(\max|(P_{1|0}-\pi_{01})/P_{1|1})-\bar{Y}_{0,1}$.

$$\begin{aligned}\Delta_{S=1}^{UB2} &= \frac{\pi_{01}}{P_{1|1}+P_{1|0}}\cdot\bar{Y}_{1,1}(\max|(P_{1|0}-\pi_{01})/P_{1|1})-\frac{2\cdot P_{1|0}-\pi_{01}}{P_{1|1}+P_{1|0}}\cdot\bar{Y}_{0,1}\\ &+ \frac{P_{1|1}+P_{1|0}-\pi_{01}}{P_{1|1}+P_{1|0}}\cdot\bar{Y}_{1,1}-\frac{P_{1|1}-P_{1|0}+\pi_{01}}{P_{1|1}+P_{1|0}}\cdot Y^{LB}\end{aligned}$$

3. $\bar{Y}_{1,1}(\max|(P_{1|0}-\pi_{01})/P_{1|1})-\bar{Y}_{0,1}(\max|(P_{1|0}-\pi_{01})/P_{1|0})$, $\bar{Y}_{1,1}(\min|(P_{1|1}-P_{1|0}+\pi_{01})/P_{1|1})-Y^{LB}$ and Δ_{01}^{UB}

$$\begin{aligned}\Delta_{S=1}^{UB3} &= \frac{P_{1|0}}{P_{1|1}+P_{1|0}}\cdot\bar{Y}_{1,1}(\max|(P_{1|0}-\pi_{01})/P_{1|1})-\frac{P_{1|0}}{P_{1|1}+P_{1|0}}\cdot\bar{Y}_{0,1}\\ &- \frac{P_{1|0}-\pi_{01}}{P_{1|1}+P_{1|0}}\cdot\bar{Y}_{0,1}(\max|(P_{1|0}-\pi_{01})/P_{1|0})\\ &+ \frac{P_{1|1}}{P_{1|1}+P_{1|0}}\cdot\bar{Y}_{1,1}-\frac{P_{1|1}-P_{1|0}+\pi_{01}}{P_{1|1}+P_{1|0}}\cdot Y^{LB}\end{aligned}$$

4. $\bar{Y}_{1,1} - \bar{Y}_{0,1}(\max|(P_{1|0} - \pi_{01})/P_{1|0})$, Δ_{10}^{UB} and Δ_{01}^{UB}

$$\begin{aligned}
\Delta_{S=1}^{UB4} &= \frac{\pi_{01}}{P_{1|1} + P_{1|0}} \cdot \bar{Y}_{1,1}(\max|(P_{1|0} - \pi_{01})/P_{1|1}) - \frac{P_{1|0}}{P_{1|1} + P_{1|0}} \cdot \bar{Y}_{0,1} \\
&- \frac{P_{1|0} - \pi_{01}}{P_{1|1} + P_{1|0}} \cdot \bar{Y}_{0,1}(\max|(P_{1|0} - \pi_{01})/P_{1|0}) + \frac{P_{1|1} + P_{1|0} - \pi_{01}}{P_{1|1} + P_{1|0}} \cdot \bar{Y}_{1,1} \\
&- \frac{P_{1|1} - P_{1|0} + \pi_{01}}{P_{1|1} + P_{1|0}} \cdot Y^{LB}
\end{aligned}$$

To show that $\Delta_{S=1}^{UB} = \Delta_{S=1}^{UB1}$ is the sharp upper bound it is sufficient to show that

$$\begin{aligned}
\frac{P_{1|0}}{P_{1|1} + P_{1|0}} \cdot \bar{Y}_{1,1}(\max|(P_{1|0} - \pi_{01})/P_{1|1}) &+ \frac{P_{1|1}}{P_{1|1} + P_{1|0}} \cdot \bar{Y}_{1,1} \geq \\
\frac{\pi_{01}}{P_{1|1} + P_{1|0}} \cdot \bar{Y}_{1,1}(\max|(P_{1|0} - \pi_{01})/P_{1|1}) &+ \frac{P_{1|1} + P_{1|0} - \pi_{01}}{P_{1|1} + P_{1|0}} \cdot \bar{Y}_{1,1}
\end{aligned}$$

and

$$\frac{2 \cdot P_{1|0} - \pi_{01}}{P_{1|1} + P_{1|0}} \cdot \bar{Y}_{0,1} \leq \frac{P_{1|0} - \pi_{01}}{P_{1|1} + P_{1|0}} \cdot \bar{Y}_{0,1}(\max|(P_{1|0} - \pi_{01})/P_{1|0}) + \frac{P_{1|0}}{P_{1|1} + P_{1|0}} \cdot \bar{Y}_{0,1}$$

Simple algebra shows that the two inequalities are always satisfied. In order to show that we do not need to maximize $\Delta_{S=1}^{UB}$ over all possible values of π_{01}, it is sufficient to see that its first derivative w.r.t. π_{01}, given by

$$\bar{Y}_{0,1} - Y^{LB} + P_{1|0} \cdot \frac{\partial \bar{Y}_{1,1}(\max|(P_{1|0} - \pi_{01})/P_{1|1})}{\partial \pi_{01}}$$

is always positive, since $\bar{Y}_{0,1} - Y^{LB} > 0$ and $P_{1|0} \cdot \frac{\partial \bar{Y}_{1,1}(\max|(P_{1|0} - \pi_{01})/P_{1|1})}{\partial \pi_{01}} > 0$. Therefore $\Delta_{S=1}^{UB}$ is maximized for $\pi_{01} = \pi_{01}^{\max}$.

A.4 Monotonicity and stochastic dominance

All bounds derived under monotonicity and stochastic dominance are special cases of the bounds derived under stochastic dominance alone given that $\pi_{01} = 0$. Therefore, the bounds are sharp.

References

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

Blanco, G., C. A. Flores, and A. Flores-Lagunes (2011): "Bounds on Quantile Treatment Effects of Job Corps on Participants' Wages," *mimeo.*

Blundell, R., A. Gosling, H. Ichimura, and C. Meghir (2007): "Changes in the Distribution of Male and Female Wages Accounting for Employment Composition Using Bounds," *Econometrica*, 75, 323–363.

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

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

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.

Gerfin, M., and M. Schellhorn (2006): "Nonparametric bounds on the effect of deductibles in health care insurance on doctor visits Swiss evidence," *Health Economics*, 15, 1011–1020.

González, L. (2005): "Nonparametric bounds on the returns to language skills," *Journal of Applied Econometrics*, 20, 771–795.

Grilli, L., and F. Mealli (2008): "Nonparametric Bounds on the Causal Effect of University Studies on Job Opportunities Using Principal Stratification," *Journal of Educational and Behavioral Statistics*, 33.

Gronau, R. (1974): "Wage comparisons-a selectivity bias," *Journal of Political Economy*, 82, 1119–1143.

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.

Horowitz, J. L., and C. F. Manski (1995): "Identification and Robustness with Contaminated and Corrupted Data," *Econometrica*, 63, 281–302.

HUBER, M. (2010): "Identification of average treatment effects in social experiments under different forms of attrition," *University of St Gallen, Dept. of Economics Discussion Paper no. 2010-22.*

HUBER, M., AND B. MELLY (2011): "Quantile regression in the presence of sample selection," *University of St Gallen, Department of Economics Discussion Paper No. 2011-09.*

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

IMBENS, G. W. (2004): "Nonparametric estimation of average treatment effects under exogeneity: a review," *The Review of Economics and Statistics*, 86, 4–29.

——— (2006): "Nonadditive models with endogenous regressors," *mimeo*, University of Chicago.

IMBENS, G. W., AND J. ANGRIST (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62, 467–475.

IMBENS, G. W., AND C. F. MANSKI (2004): "Confidence Intervals for Partially Identified Parameters," *Econometrica*, 72(6), 1845–1857.

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

KREIDER, B., AND S. C. HILL (2009): "Partially Identifying Treatment Effects with an Application to Covering the Uninsured," *Journal of Human Resources*, 44, 409–449.

LECHNER, M., AND B. MELLY (2007): "Earnings Effects of Training Programs," *IZA Discussion Paper no. 2926.*

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

MANSKI, C. F. (1989): "Anatomy of the selection problem," *The Journal of Human Resources*, 24, 343–360.

——— (1994): "The selection problem," in *Advances in Econometrics: Sixth World Congress*, ed. by C. Sims., pp. 143–170. Cambridge University Press.

——— (1997): "Monotone Treatment Response," *Econometrica*, 65, 1311–1334.

MEALLI, F., AND B. PACINI (2008a): "Causal inference with nonignorably missing outcomes: instrumental variables and principal stratification," *mimeo.*

——— (2008b): "Comparing principal stratification and selection models in parametric causal inference with nonignorable missingness," *Computational Statistics & Data Analysis*, 53(2), 507–516.

MELLACE, G., AND R. ROCCI (2011): "Principal Stratification in sample selection problems with non normal error terms," CEIS Research Paper 194, Tor Vergata University, CEIS.

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

——— (2009): “Two-step series estimation of sample selection models,” *Econometrics Journal*, 12, S217–S229.

Newey, W. K., and D. McFadden (1994): “Large Sample Estimation and Hypothesis Testing,” in *Handbook of Econometrics*, ed. by R. Engle, and D. McFadden. Elsevier, Amsterdam.

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

Robins, J. (1989): “The Analysis of Randomized and Non-RandomizedAIDS Treatment Trials Using a New Approach to Causal Inference in Longitudinal Studies,” in *Health Service Research Methodology: A Focus on AIDS*, ed. by L. Sechrest, H. Freeman, and A. Mulley, pp. 113–159. U.S. Public Health Service, Washington, DC.

Rubin, D. B. (1977): “Assignment to treatment group on the basis of a covariate,” *Journal of Educational Statistics*, 2, 1–26.

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

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.

——— (2009): “Likelihood-Based Analysis of Causal Effects of Job-Training Programs Using Principal Stratification,” *Journal of the American Statistical Association*, 104(485), 166–176.