Christelis, Dimitris; Messina, Julián

Working Paper

Partial identification of population average and quantile treatment effects in observational data under sample selection

IDB Working Paper Series, No. IDB-WP-985

Provided in Cooperation with:
Inter-American Development Bank (IDB), Washington, DC

Suggested Citation: Christelis, Dimitris; Messina, Julián (2019) : Partial identification of population average and quantile treatment effects in observational data under sample selection, IDB Working Paper Series, No. IDB-WP-985, Inter-American Development Bank (IDB), Washington, DC, https://doi.org/10.18235/0001596

This Version is available at:
https://hdl.handle.net/10419/208173

IDB WORKING PAPER SERIES Nº IDB-WP-985

Partial Identification of Population Average and Quantile Treatment Effects in Observational Data under Sample Selection

Dimitris Christelis
Julián Messina

Inter-American Development Bank
Department of Research and Chief Economist

March 2019

Inter-American Development Bank

Partial Identification of Population Average and Quantile Treatment Effects in Observational Data under Sample Selection

Dimitris Christelis*
Julián Messina**

* University of Naples Federico II, CSEF, CFS, CEPAR and Netspar
** Inter-American Development Bank and IZA

March 2019

Cataloging-in-Publication data provided by the
Inter-American Development Bank
Felipe Herrera Library

Christelis, Dimitris.
Partial identification of population average and quantile treatment effects in observational data under sample selection / Dimitris Christelis, Julián Messina.
p. cm. — (IDB Working Paper Series ; 985)
Includes bibliographic references.
1. Mathematical ability-Testing-Brazil-Econometric models. 2. Sampling (Statistics).
I. Messina, Julián, 1971- II. Inter-American Development Bank. Department of Research and Chief Economist. III. Title. IV. Series.
IDB-WP-985

http://www.iadb.org

Abstract[*]

This paper partially identifies population treatment effects in observational data under sample selection, without the benefit of random treatment assignment. Bounds are provided for both average and quantile population treatment effects, combining assumptions for the selected and the non-selected subsamples. We show how different assumptions help narrow identification regions, and we illustrate our methods by partially identifying the effect of maternal education on the 2015 PISA math test scores in Brazil. We find that while sample selection increases considerably the uncertainty around the effect of maternal education, it is still possible to calculate informative identification regions.

JEL classifications: C21, C24, I2
Keywords: Sample selection, Population treatment effects, Partial identification, Bounds, Observational data, PISA, Brazil

[*] We thank Diana Hincapié, Charles Manski, Oscar Mitnik, and seminar participants at the CSEF-IGIER conference, University of Roma Tor Vergata and the Inter-American Development Bank. All errors are our own.

1. Introduction

In observational data, the identification of causal effects is considerably complicated by treatment selection, that is, by the fact that sample units are observed having only one among all possible values of the treatment. If the treatment is not randomly assigned, or a quasi-experimental situation does not obtain, then it is difficult to justify projecting outcomes observed in a subsample receiving a particular treatment value to subsamples observed receiving different values.

Sample selection makes the identification of causal effects even more difficult in observational data because it adds to the problem of unobserved outcomes due to treatment selection the problem of unobserved outcomes and possibly unobserved treatments in the non-selected subsample. As sample selection is often non-random (e.g., it can be due to a decision), ignoring it can lead to biased estimates of treatment effects in the population.

There are many situations in which researchers may be interested in treatment effects in the whole population, and not just in the selected subsample. This is especially true when the outcome is an important indicator of the state of the population, or when the size and composition of the selected subsample can change over time due to circumstances or policy changes.

Labor economists may want to estimate the effect of education on earnings even for those not currently working, as the latter are still part of the labor force (if unemployed) or can be so in the future. If unemployment is involuntary, this is arguably the appropriate measure to compute the returns to schooling (Ashenfelter and Ham, 1979). In finance, researchers may want to know the effect of a change in capital gains taxation on the share of portfolios invested in stocks not only among actual stockowners but also among potential stock investors who take capital gains taxation into consideration when making their investment decisions. Trade economists may want to know if some policy intervention would affect exports not only among exporting firms, but also among potential exporters who are currently not engaged in trade.

Sample selection can be a problem not only when analyzing observational data, but also data from randomized control trials. For example, randomized control trials often include subjects who volunteer to participate in them, who may form a non-random sample of the population of interest. It is also standard to exclude from a medical trial persons who suffer from multiple diseases, which can also generate a form of sample selection.

In this paper, we use partial identification (PI) methods to derive bounds on the average and quantile treatment effects (ATEs and QTEs, respectively) for the whole population under

sample selection in observational data, when the outcome is never observed in the non-selected subsample, while there are no restrictions on how often the treatment is observed in that subsample. We calculate identification regions of the treatment effect under various assumptions that are weaker, and thus more credible, than those used in the literature to achieve point identification of treatment effects (see, e.g., Gronau, 1973; Heckman, 1974, 1976, 1979; Lee, 1982; Das, Newey and Vella, 2003; Newey, 2009). Our approach can be viewed as providing a general framework for partially identifying treatment effects in observational data, with the cases of no sample selection discussed in Manski (1994, 1997) and Giustinelli (2011) being special ones.

We apply our results to the estimation of the causal effect of maternal education on children math test scores in Brazil using data from the 2015 edition of the Programme for International Student Assessment (PISA; see OECD, 2017a). The estimation of the causal impact of parental education on children's learning has been extensively studied in the literature (for a recent survey see Holmlund, Lindahl and Plug, 2011). The identification of this causal effect is hindered by the fact that children whose parents have different levels of education are likely to be systematically different in many other respects (e.g., socio-economic status, health, family networks), and these systematic differences make it hard to disentangle the impact of parental education on school performance. To address this problem, researchers have used twin studies, differences in school outcomes between biological and adopted children, and instrumental variable (IV) estimation, with varying degrees of success.

In our context, an additional obstacle to identification arises from sample selection due to children that: i) drop out of school before the completion of the legally compulsory education level, and in any case before age 15, which is the age at which the PISA math assessment is administered in the schools; or ii) lag behind in school badly enough not to be included in the PISA sampling frame, i.e., they attend school but are below grade 6 at age 15. This sample selection makes the subsample of children that take the PISA test likely not representative, as dropping out and lagging behind are unlikely to be random.

Using PI under weak assumptions, we provide identification regions for the effect of maternal education on math scores for the whole population of children at age 15, that is, irrespective of whether they actually attend school at grade 6 or above or not. Hence, this magnitude also incorporates the effect of maternal education on the decision to drop out and on lagging behind in school. We want to know what happens to the human capital of the whole

population of 15-year old children (as measured by the test scores) because children who have dropped out of school or lag behind by the time they are 15 years old are also going to be workers, entrepreneurs, and more generally decision makers in adulthood (or even earlier than that). Importantly, their human capital is going to be a crucial determinant of their productivity as well as of the quality of their decisions.

Our paper is related to various strands of the PI literature applied to observational data. In the absence of sample selection, Manski (1990, 1997), and Manski and Pepper (2000) partially identify ATEs, while Manski (1994) and Giustinelli (2011) partially identify QTEs. Manski (1989) and Blundell et al. (2007) provide identification regions for population descriptive statistics under sample selection. Molinari (2012) discusses identification when information on treatments is partially missing, while outcomes are always observed. Finally, De Haan (2011) estimates bounds for the effect of parental education on offspring years of schooling, abstracting from sample selection.

There are also several papers that bound treatment effects in the context of randomized control trials. Horowitz and Manski (2000) bound the ATE under sample selection using a binary outcome, while Lee (2009) and Blanco, Flores and Flores-Lagunes (2012) bound it under sample selection for those always selected. Huber and Mellace (2015) bound ATEs under sample selection on the subsamples of compliers, defiers, and those whose outcomes are observed. Huber, Laffers and Mellace (2017) bound the ATE on the treated and other subpopulations using an exogenous IV. Chen, Flores and Flores-Lagunes (2018) bound the population ATE under no sample selection also using an exogenous IV. We, on the other hand, use observational data under sample selection, and thus our calculations are complicated by the lack of random treatment assignment. Moreover, we bound both ATE and QTEs for the whole population.

2. Methodology

2.1 Setup of the Problem

Following Manski (1997), we assume that for each individual i there is a response function $y_i(\bullet): D \to Y$ that maps mutually exclusive and exhaustive treatments $d \in D$ into outcomes $y_i(d) \in Y$. These response functions $y_i(\bullet)$ can differ across individuals in arbitrary ways. We posit a common infimum and supremum for all $y_i(\bullet)$, denoted by y^{INF} and y^{SUP}, respectively. Let w_i denote the realized treatment (mother's education in our case) received by i, and $y_i \equiv y_i(w_i)$ the

associated observed outcome, which in our context is the PISA math test score. We denote by $y_i(d)$ a latent potential outcome when $w_i \neq d$, and by $G(w)$ the realized treatment distribution.

As regards sample selection, let S be a binary indicator that is equal to one if the individual is part of the selected subsample (that is, takes the PISA test in our context), and equal to zero if not.

Under sample selection, the data allow us to identify the probability distribution $H(S)$ of the selection indicator as well as the joint distribution of outcomes and realized treatments in the selected subsample $T(y, w|S = 1)$.[1] We make no assumptions about the observability of the distribution of the realized treatment in the non-selected subsample $G(w|S = 0)$. Our aim is to identify the distribution of the response function $F[y(d)]$ so as to estimate the ATEs and QTEs in the population.

The population ATEs and QTEs record the differences in outcomes at the mean and at different quantiles in a particular kind of non-randomized experiment, namely one in which all population units take two different treatment values. This is clearly a counterfactual setup, as units in the selected subsample are actually observed taking only one treatment value at any given point in time. Non-selected units as well can take only one treatment value at any given point in time and, in addition, their outcomes are never observed. Hence, population outcomes are only partially observed.

This setup has also some desirable features. First, the whole population is assumed to take two different treatment values; hence, control and treated groups coincide with the population. Thus, the problem of systematic differences not due to different treatment values between the control and treated groups is ruled out by construction (the distribution of all variables other than the outcome and the treatment is taken as given). Second, as this counterfactual non-randomized experiment takes place in a large sample that is representative of the population, external validity of results is less of a concern.

[1] All results go through if we condition additionally on a vector of controls X; thus, we omit such conditioning to economize on notation.

2.1.1 Average Treatment Effects

By the law of total probability, the mean potential outcome $E[y(d)]$ is equal to

$$\begin{aligned} E[y(d)] = E[y(d)|S=1]P(S=1) + E[y(d)|S=0]P(S=0) = \\ \{E[y|w=d,S=1]P(w=d|S=1) + \\ E[y(d)|w\neq d,S=1]P(w\neq d|S=1)\}P(S=1) \ + \\ E[y(d)|S=0]P(S=0). \end{aligned} \tag{1}$$

There are two unobserved terms in (1), namely the counterfactual mean outcome $E[y(d)|w\neq d,S=1]$ in the selected subsample, and the unobserved mean outcome $E[y(d)|S=0]$ in the non-selected subsample. In the first case, selection into treatment and sample selection imply that $E[y(d)|w\neq d,S=1]\neq E[y|w=d,S=1]$, and thus it is not possible to substitute the latter term for the former. In the second case, there is no obvious candidate to put in place of $E[y(d)|S=0]$ without imposing further assumptions on the non-selected subsample.

The solution to the problem of estimating counterfactual and unobserved outcomes proposed by Manski (1989, 1990) is to bound them from above and below, thus also bounding $E[y(d)]$, which becomes partially identified. In our case, one can put a lower and upper bound on $E[y(d)|S=1]$ using a set of assumptions L (denoted by $SLB_E^L(d)$ and $SUB_E^L(d)$, respectively), independently from putting, using a set of assumptions M, a lower and upper bound on $E[y(d)|S=0]$ (denoted by $NLB_E^M(d)$ and $NUB_E^M(d)$, respectively). Then one can bound $E[y(d)]$ as follows:

$$\begin{aligned} LB_E^{L,M}(d) = SLB_E^L(d)P(S=1) + NLB_E^M(d)P(S=0) \\ \leq E[y(d)] \leq \\ UB_E^{L,M}(d) = SUB_E^L(d)P(S=1) + NUB_E^M(d)P(S=0). \end{aligned} \tag{2}$$

Equations (1) and (2) show that one can use sample selection to separate the whole sample into two subsamples, the selected and the non-selected. We keep those two subsamples fixed throughout all our analyses and examine what happens to mean potential outcomes separately in each subsample. This in turn allows us to focus on terms that are not observed in each subsample independently of what happens to the other subsample and calculate the associated bounds. The population result at the mean is always a weighted average of the results in the two fixed subsamples, with the weights being the probability of selection and its complement. In other words,

observed selection outcomes S are given (as is the case with actual treatments received w), and we use this fact to keep the conditioning in (1) and (2) constant throughout our analyses.

Using a sample split by a given observed selection outcome does not imply, however, that selection does not depend on maternal education. We also note that the observed selection outcome is a different concept from the potential selection outcome (denoted by $l(d)$). Unobserved potential selection outcomes $l(d)$ can differ from observed ones just as unobserved potential outcomes $y(d)$ can differ from realized ones. For example, selected children could have been non-selected had their maternal education been lower than the actual one, while non-selected children could have become selected ones had their maternal education been higher than the actual one.

As we discuss below, when introducing assumptions about counterfactual outcomes $y(d)$ in both subsamples, we take into account how different (and potentially counterfactual) treatment values can affect $y(d)$ through their effect on $l(d)$. We stress, however, that this analysis is always carried out within each fixed subsample defined by the observed selection outcome S. For example, we can examine separately for the observed selected and non-selected children how lower maternal education can affect the probability of school attendance at grade 6 or above at age 15, and through it the associated possibly counterfactual test score.

The ATE of a change in the treatment from d_1 to d_2 is defined as

$$ATE(d_2, d_1) = E[y(d_2)] - E[y(d_1)]. \tag{3}$$

Using (2) and (3), the ATE(d_2, d_1) can be bounded from below and above as follows:

$$LB_E^{L,M}(d_2) - UB_E^{L,M}(d_1) \leq ATE(d_2, d_1) \leq UB_E^{L,M}(d_2) - LB_E^{L,M}(d_1). \tag{4}$$

2.1.2 Quantile Treatment Effects

Turning now to QTEs, we first note that the cumulative distribution of the response function $F[y(d)]$ is equal to

$$\begin{aligned} F[y(d)] = F[y(d)|S=1]P(S=1) + F[y(d)|S=0]P(S=0) = \\ \{F[y|w=d, S=1]P(w=d|S=1) + \\ F[y(d)|w \neq d, S=1]P(w \neq d|S=1)\}P(S=1) + \\ F[y(d)|S=0]P(S=0). \end{aligned} \tag{5}$$

As with $E[y(d)]$, we have a counterfactual term due to treatment selection $F[y(d)|w \neq d, S = 1]$ and an unobserved term $F[y(d)|S = 0]$ due to sample selection. We denote the lower and upper bound on $F[y(d)|S = 1]$ using a set of assumptions L as $SLB_F^L(d)$ and $SUB_F^L(d)$, respectively. Moreover, we denote the lower upper bound on $F[y(d)|S = 0]$ using a set of assumptions M (denoted by $NLB_F^M(d)$ and $NUB_F^M(d)$, respectively). Then one can bound $F[y(d)]$ as follows:

$$\begin{aligned} LB_F^{L,M}(d) = SLB_F^L(d)P(S = 1) + NLB_F^M(d)P(S = 0) \\ \leq F[y(d)] \leq \\ UB_F^{L,M}(d) = SUB_F^L(d)P(S = 1) + NUB_F^M(d)P(S = 0). \end{aligned} \tag{6}$$

Once more, the lower and upper bounds on $F[y(d)]$ are equal to the weighted average of the respective bounds in the two subsamples.

As Blundell et al. (2007) point out, to get the lower bound $LB_\alpha^{L,M}(d)$ on the α-quantile of $F[y(d)]$ (denoted by $Q_\alpha[y(d)]$) one needs to invert the upper bound $UB_F^{L,M}(d)$ in (6), that is, $LB_\alpha^{L,M}(d)$ is equal to the value of y that solves the equation $UB_F^{L,M}(d) = \alpha$. Correspondingly, the upper bound $UB_\alpha^{L,M}(d)$ on $Q_\alpha[y(d)]$ is equal to the value of y that solves the equation $LB_F^{L,M}(d) = \alpha$. We note that neither the lower bound $LB_\alpha^{L,M}(d)$ nor the upper bound $UB_\alpha^{L,M}(d)$ on $Q_\alpha[y(d)]$ is in general equal to the weighted average (with the weights equal to the probability of selection and its complement) of the corresponding bounds on the α-quantile in the selected and non-selected subsamples, as inverting the bounds on $F[y(d)]$ is a nonlinear operation.

The α-QTE of a change in D from d_1 to d_2 is defined as

$$\alpha\text{-}QTE(d_2, d_1) = Q_\alpha[y(d_2)] - Q_\alpha[y(d_1)], \tag{7}$$

where $Q_\alpha[y(d)]$ denotes the α-quantile of the distribution of the response function $F[y(d)]$. Using (7), the α-QTE(d_2, d_1) can be bounded from below and above as follows:

$$LB_\alpha^{L,M}(d_2) - UB_\alpha^{L,M}(d_1) \leq \alpha\text{-QTE}(d_2, d_1) \leq UB_\alpha^{L,M}(d_2) - LB_\alpha^{L,M}(d_1). \tag{8}$$

2.2 No Assumption Bounds[2]

Starting with mean potential outcomes, if one is unwilling to make any assumptions about the bounds on the counterfactual and unobserved terms, then the most conservative solution is to put them equal to the infimum and supremum of $y(d)$, denoted by y^{INF} and y^{SUP}, respectively. To make the no assumption (NA) bounds operational, we use two values for y^{INF}. The first value, denoted by y^{INFS}, is equal to the observed minimum value of outcome (i.e., the PISA test score in our context) and is applied to the selected subsample. We put the second value, denoted by y^{INFN}, equal to zero, as we cannot rule out the possibility that children who do not actually take the PISA test would have had such a low score had they taken it. Correspondingly, there can be two different values for y^{SUP}, namely y^{SUPS} for the selected and y^{SUPN} for the non-selected subsample, and there is little reason to rank them in our context, as the best the non-selected children could have done in the test might be better or worse than the best actual test takers can do. In practice, we put both y^{SUPS} and y^{SUPN} equal to the observed maximum test score. We consider the choice of the observed extrema to be conservative, as they bound counterfactual or unobserved mean outcomes rather than individual outcomes.

Given the decomposition of the unobserved mean potential outcome in (1), applying NA bounds to the selected and non-selected subsamples (denoted by NA,NA) bounds $E[y(d)]$ as follows:

$$\begin{aligned}&\{E[y|w=d,S=1]P(w=d|S=1)+y^{INFS}P(w\neq d|S=1)\}P(S=1)\\&\quad+y^{INFN}P(S=0)\\&\qquad\qquad\leq E[y(d)]\leq\\&\{E[y|w=d,S=1]P(w=d|S=1)+y^{SUPS}P(w\neq d|S=1)\}P(S=1)\\&\quad+y^{SUPN}P(S=0).\end{aligned}\tag{9}$$

In the absence of any further assumptions, the bounds in (9) cannot be improved upon, that is, they are sharp. Moreover, if $y^{INFN}\leq y^{INFS}<E[y|w=d,S=1]<y^{SUPS}\leq y^{SUPN}$, then (9) implies that the NA,NA lower (higher) bound under sample selection is smaller (larger) to the NA,NA bound under no sample selection (obtained by putting $P(S=1)=1$ in (9)). This is to be expected, as sample selection creates more uncertainty about $E[y(d)]$.

[2] In the related literature, these bounds are also called worst case bounds, or bounds using no information or using only empirical evidence (see, e.g., Manski, 1997; Giustinelli, 2011).

Turning to the NA,NA bounds on $F[y(d)]$, we use, as in Manski (1989: 346), the fact that a lower (upper) bound on probabilities (and thus cumulative distributions) that entails no assumptions is zero (one). Hence, and given (5), the NA,NA bounds on $F[y(d)]$ can be expressed as follows:

$$\begin{gathered} F[y|w=d,S=1]P(w=d|S=1)P(S=1) \\ \leq F[y(d)] \leq \\ \{F[y|w=d,S=1]P(w=d|S=1)+P(w\neq d|S=1)\}P(S=1)+P(S=0). \end{gathered} \tag{10}$$

In the absence of any further assumptions, the bounds in (10) are sharp. The bounds on the α-quantile are described in the following proposition.

Proposition 1 *Let* $\alpha \in (0,1)$. *Define* $r^{NA,NA}(\alpha,d)$ *and* $s^{NA,NA}(\alpha,d)$ *as*

$$r^{NA,NA}(\alpha,d) = \begin{cases} Q_{\left[1-\frac{(1-\alpha)}{P(w=d|S=1)P(S=1)}\right]}(y|w=d,S=1) \text{ if } P(w\neq d|S=1)P(S=1)+ \\ \qquad\qquad P(S=0)<\alpha<1, \\ min(\,y^{INFS},\,y^{INFN}) \qquad otherwise, \end{cases} \tag{11}$$

$$s^{NA,NA}(\alpha,d) = \begin{cases} Q_{\left[\frac{\alpha}{P(w=d|S=1)P(S=1)}\right]}(y|w=d,S=1) \text{ if } 0<\alpha\leq P(w=d|S=1)P(S=1), \\ max(\,y^{SUPS},\,y^{SUPN}) \qquad otherwise. \end{cases}$$

Then, $\forall\, d \in D$,

$$r^{NA,NA}(\alpha,d) \leq Q_\alpha[y(d)] \leq s^{NA,NA}(\alpha,d). \tag{12}$$

In the absence of other information, these bounds are sharp.

Proof. See Appendix A.1.

The NA,NA bounds in (11) become equal to the NA bounds in Manski (1994) when there is no sample selection problem, that is, when $P(S=1)=1$. Thus, our approach can be viewed as providing a general framework for partially identifying treatment effects in observational data, with the case of no sample selection being a special one.

As is clear from the results in Manski (1994) for the case of no sample selection, the NA,NA bounds on $Q_\alpha[y(d)]$ in (11) are more likely to be uninformative compared to the NA

bounds under no sample selection. Furthermore, even when the NA,NA bounds in (11) are informative, they produce wider identification regions for $Q_\alpha[y(d)]$ than in the absence of sample selection, as is the case with $E[y(d)]$ discussed above.

We note that the NA,NA bounds assume nothing about the distribution $G(w|S=0)$ of the treatment variable in the non-selected subsample. This also implies that the treatment variable can be always, partially or never observable in that subsample. Even if it were always observable, however, the fact that the outcome is never observable makes it impossible to use $G(w)$ to construct bounds on magnitudes defined in the non-selected subsample without any further assumptions.

2.3 Assumptions on the Selected Subsample

To further narrow the identification region of the ATE and the α-QTE we use additional assumptions that apply to the selected subsample (i.e., the students who take the PISA math test).

2.3.1 Monotone Treatment Response

The monotone treatment response (MTR) was introduced by Manski (1997), and it posits that a higher level of the treatment does not reduce the outcome. In Manski (1997) the assumption is formulated as holding for every sampling unit within the selected subsample, that is, $\forall i$ and $d_1, d_2 \in D$ such that $d_2 > d_1$ and for $S=1$

$$y_i(d_2) \geq y_i(d_1). \tag{13}$$

As discussed in Appendix A.2, one can obtain all results related to the MTR assumption by using instead of (13) the weaker assumption that distributions of outcomes under higher treatment levels dominate distributions under lower treatment levels. This weak stochastic dominance holds in all subsamples of the selected subsample that are defined by a particular level of the treatment, and thus it is not necessary that MTR hold for every selected unit, as in (13).

Formally, the MTR assumption states that $\forall\, d, d_1, d_2 \in D$ such that $d_2 > d_1$, $F[y(d_2)|w=d, S=1] \geq_d F[y(d_1)|w=d, S=1]$, that is,

$$F[y(d_2)|w=d, S=1] \leq F[y(d_1)|w=d, S=1], \tag{14}$$

where $\geq_d$ denotes weak stochastic dominance. The MTR assumption is often a mild one. For example, it is reasonable to use it when studying the effect of education and work experience on wages, or the effect of physical exercise on life expectancy. It is also a mild assumption in our case, as it is unlikely that a higher level of maternal education harms children's school performance, and it's even more unlikely that this happens over a distribution of outcomes in violation of (14). Both (13) and (14) are, however, untestable assumptions because they involve counterfactual comparisons.

Another channel through which (14) is likely to hold is that a higher maternal education should weakly increase the probability of school attendance at grade 6 or above at age 15, that is $E[l(d_2)|w = d, S = 1] \geq E[l(d_1)|w = d, S = 1]$. This in turn should lead to weakly higher scores.

Given the general formulation of the bounds on $E[y(d)]$ in (2), the lower (upper) bound on $E[y(d)]$, when using MTR on the selected subsample and NA on the non-selected one, is a weighted average of the MTR lower (upper) bound under no sample selection SLB_E^{MTR} (SUB_E^{MTR}) derived in Manski (1997), and of y^{INFN} (y^{SUPN}), the weights being equal to the probability of selection and its complement, that is,

$$\begin{gathered}\{E[y|w \leq d, S = 1]P(w \leq d|S = 1) + y^{INFS}P(w > d|S = 1)\}P(S = 1) + y^{INFN}P(S = 0 \\ \leq E[y(d)] \leq \\ \{y^{SUPS}P(w < d|S = 1) + E[y|w \geq d, S = 1]P(w \geq d|S = 1)\}P(S = 1) + \\ y^{SUPN}P(S = 0).\end{gathered} \tag{15}$$

As Manski (1997) shows, the MTR bounds on $E[y(d)|S = 1]$ are sharp, and thus the bounds in (15) are also sharp in the absence of any further assumptions. Moreover, if $y^{INFN} \leq y^{INFS} < E[y|w \leq d, S = 1]$ and $E[y|w \geq d, S = 1] < y^{SUPS} \leq y^{SUPN}$, then the identification regions in (15) are wider than those under no sample selection derived in Manski (1997) due to the conservative bounds y^{INFN} and y^{SUPN} used in the non-selected subsample.

Unlike the MTR,NA bounds on $E[y(d)]$, the MTR,NA bounds on $Q_\alpha[y(d)]$ cannot be computed as weighted averages of the bounds in the selected and non-selected subsample. Rather, they are described in the following proposition.

Proposition 2 *Let the MTR assumption in* (14) *hold. Let* $\alpha \in (0,1)$. *Define* $r^{MTR,NA}(\alpha, d)$ *and* $s^{MTR,NA}(\alpha, d)$ *as*

$$
\begin{aligned}
& r^{MTR,NA}(\alpha, d) \\
&= \begin{cases} Q_{\left[1-\frac{(1-\alpha)}{P(w\le d|S=1)P(S=1)}\right]}(y|w \le d, S=1) \text{ if} \\ \qquad P(w>d|S=1)P(S=1) + (S=0) < \alpha < 1, \\ min(\, y^{INFS},\, y^{INFN}) \qquad otherwise, \end{cases} \\
& s^{MTR,NA}(\alpha, d) \\
&= \begin{cases} Q_{\left[\frac{(1-\alpha)}{P(w\ge d|S=1)P(S=1)}\right]}(y|w \ge d, S=1) \text{ if} \\ \qquad 0 < \alpha \le P(w \ge d|S=1)P(S=1), \\ max(\, y^{SUPS},\, y^{SUPN}) \qquad otherwise. \end{cases}
\end{aligned}
\tag{16}
$$

Then, $\forall\, d \in D$,

$$
r^{MTR,NA}(\alpha, d) \le Q_\alpha[y(d)] \le s^{MTR,NA}(\alpha, d). \tag{17}
$$

In the absence of other information, these bounds are sharp.

Proof. See Appendix A.2.

Once again, the bounds in (16) become equal to the MTR bounds in Manski (1997) when there is no sample selection, that is, when $P(S=1)=1$.

As is the case with the NA bounds, sample selection increases the probability that the MTR,NA bounds on $Q_\alpha[y(d)]$ in (16) are uninformative compared to the MTR bounds under no sample selection derived in Manski (1997). Furthermore, even when MTR,NA bounds are informative, they produce wider identification regions for $Q_\alpha[y(d)]$ compared to the case of no sample selection.

We note that imposing the MTR assumption on the selected subsample does not ensure that the lower bound of the ATE and α-QTE is equal to zero, as is the case in the absence of sample selection (see Manski, 1997; Giustinelli, 2011). This is so because the bounds on the non-selected subsample are still those under NA, that is, equal to y^{INFN} and y^{SUP} and thus very conservative. As a result, the identification regions after imposing MTR only on the selected subsample can still be wide enough to make the lower bound of the ATE and α-QTE negative. As discussed in Section 2.4.1 below, only when one imposes MTR also on the non-selected subsample it is assured that the lower bound of the ATE and α-QTE cannot be lower than zero.

2.3.2 Monotone Treatment Selection

Monotone Treatment Selection (MTS) was introduced by Manski and Pepper (2000, henceforth MP) for mean outcomes, and by Giustinelli (2011) for quantiles. In our application, MTS implies that children who attend school at grade 6 or above and whose mothers have higher education would do on average at least as well as children whose mothers have low education if both children groups had, counterfactually, mothers with the same level of education. This is likely a mild assumption, as a higher observed maternal education implies that the child likely has a number of advantages in life that help him/her under any circumstances (e.g., higher level of economic resources, access to better peers and a safer environment from living in a better neighborhood).

While MP formulate the MTS assumption in terms of means, and Giustinelli (2011) in terms of inequalities of quantiles, we formulate the assumption in terms of stochastic dominance of outcome distributions. Given that both means and quantiles respect stochastic dominance, this formulation encompasses both the formulation of MP and that of Giustinelli (2011).

Formally, the MTS assumption states that $\forall\, d, d_1, d_2 \in D$ such that $d_2 > d_1$, $F[y(d)|w = d_2, S = 1] \geq_d F[y(d)|w = d_1, S = 1]$, that is,

$$F[y(d)|w = d_2, S = 1] \leq F[y(d)|w = d_1, S = 1]. \tag{18}$$

One can think about the MTS assumption as a particular form of non-random selection into treatment, that is, a particular form of the condition $F[y(d)|w \neq d, S = 1] \neq F[y|w = d, S = 1]$, which implies that those who choose different levels of the treatment have systematically different outcomes also under a counterfactual common treatment. The MTS assumption pins down the direction of this difference, as it states that higher observed treatment levels lead to weakly dominating distributions of potential outcomes under a counterfactual common treatment value.

As was the case with MTR, another channel through which MTS is likely to hold is that a higher observed maternal education should weakly increase the probability of school attendance at grade 6 at age 15, even if the counterfactual maternal education received is the same. In other words, $E[l(d)|w = d_2, S = 1] \geq E[l(d)|w = d_1, S = 1]$. The reasoning for this is the same as with the potential scores, that is, a higher observed socio-economic status should weakly help with school attendance and progress.

By (2), the lower (upper) bound on $E[y(d)]$, when using MTS on the selected subsample and NA on the non-selected one, is a weighted average of SLB_E^{MTS} (SUB_E^{MTS}), namely the MTS lower (upper) bound under no sample selection derived in MP, and of y^{INFN} (y^{SUPN}), that is,

$$\begin{aligned}&\{y^{INFS}P(w<d|S=1)+E[y|w=d,S=1]P(w\geq d|S=1)\}P(S=1)\\&\quad+y^{INFN}P(S=0)\\&\qquad\leq E[y(d)]\leq\\&\{E[y|w=d,S=1]P(w\leq d|S=1)+y^{SUPS}P(w>d|S=1)\}P(S=1)\\&\quad+y^{SUPN}P(S=0).\end{aligned}\tag{19}$$

MP show that the MTS bounds SLB_E^{MTS} and SUB_E^{MTS} in (19) are sharp, and thus the bounds in (19) are also sharp in the absence of any further assumptions. Moreover, if $y^{INFN}\leq y^{INFS}<E[y|w=d,S=1]<y^{SUPS}\leq y^{SUPN}$, then the identification regions in (19) are wider than those under no sample selection derived by MP due to the conservative bounds y^{INFN} and y^{SUPN} used in the non-selected subsample.

Once more, the MTS,NA bounds on $Q_\alpha[y(d)]$ cannot be computed as simple weighted averages of the bounds in the selected and non-selected subsample, unlike the MTS,NA bounds on $E[y(d)]$ in (19). Rather, they are described in the following proposition.

Proposition 3 *Let the MTS assumption* (18) *hold. Let* $\alpha\in(0,1)$. *Define* $r^{MTS,NA}(\alpha,d)$ *and* $s^{MTS,NA}(\alpha,d)$ *as*

$$\begin{aligned}&r^{MTS,NA}(\alpha,d)\\&=\begin{cases}Q_{\left[1-\frac{(1-\alpha)}{P(w\geq d|S=1)P(S=1)}\right]}(y|w=d,S=1) & \text{if } P(w<d|S=1)P(S=1)+\\ & \qquad (S=0)<\alpha<1,\\ min(\,y^{INFS},\,y^{INFN}) & otherwise,\end{cases}\\&s^{MTS,NA}(\alpha,d)=\\&\begin{cases}Q_{\left[\frac{\alpha}{P(w\leq d|S=1)P(S=1)}\right]}(y|w=d,S=1) \text{ if} & \\ \qquad\qquad \alpha\leq P(w\leq d|S=1)P(S=1),\\ max(\,y^{SUPS},\,y^{SUPN}) \qquad otherwise.\end{cases}\end{aligned}\tag{20}$$

Then, $\forall\ d\in D$,

$$r^{MTS}(\alpha,d)\leq Q_\alpha[y(d)]\leq s^{MTS}(\alpha,d).\tag{21}$$

In the absence of other information, these bounds are sharp.

Proof. See Appendix A.3.

Once again, when there is no sample selection, that is when $P(S=1)=1$, the bounds in (20) become equal to the MTS bounds under no sample selection derived in Giustinelli (2011).

As is the case with the NA,NA and MTR,NA bounds, sample selection increases the probability that the MTS,NA bounds on $Q_\alpha[y(d)]$ in (20) are uninformative compared to the MTS bounds under no sample selection derived in Giustinelli (2011). Furthermore, even when the MTS,NA bounds under sample selection are informative, they produce wider identification regions for $Q_\alpha[y(d)]$ than under no sample selection.

The MTS assumption is, like the MTR one, counterfactual, because it compares outcomes between subsamples actually receiving different treatments under the counterfactual situation in which they received the same treatment. The combination of MTR and MTS, however, is testable, as shown in MP (p. 1004, footnote 9). In particular, MP show that MTR and MTS jointly imply that $\forall\, d_1, d_2 \in D$ such that $d_2 > d_1$,

$$E(y|w=d_2, S=1) \geq E(y|w=d_1, S=1). \tag{22}$$

Equation (22) states that the MTR and MTS assumptions jointly imply that the observed mean outcomes are weakly increasing in the value of the treatment. It is easy to show that an analogous result applies also to quantiles, as the analysis of the implications of the joint MTR and MTS assumption can be easily expressed in terms of potential outcome distributions. As we discuss in Section 3 below, in our data we observe a very clear positive association between the outcome and the treatment, and thus we cannot reject the joint MTR and MTS assumption.

By combining the MTR and MTS assumptions it is possible to compute identification regions that are typically narrower than those derived using either of the two assumptions. As Giustinelli (2011) points out, the combination of MTR and MTS produces bounds on $F[y(d)|S=1]$ that are linear combinations of observed cumulative distributions, and thus cannot be inverted analytically to derive bounds on $Q_\alpha[y(d)|S=1]$. As discussed in Appendix A.3, this is true also for the MTR+MTS,NA bounds on $Q_\alpha[y(d)]$. Hence, these bounds are calculated by numerical inversion of the bounds on $F[y(d)]$. On the other hand, we show in Appendix A.3 that MTR+MTS,NA bounds on $E[y(d)]$ can be computed analytically as weighted averages of the bounds in the selected and non-selected subsamples.

2.4 Assumptions on the Non-Selected Subsample

One can further narrow the identification region of the ATE and the α-QTE by imposing assumptions on the non-selected subsample (i.e., the students who have dropped out of school or lag behind). In particular we discuss the MTR, stochastic dominance and bounded variation assumptions. As will be clear from the discussion below, all three assumptions refer to distributions of the potential outcome $y(d)$, either in the whole sample or in the selected and the non-selected subsamples. As a result, none of these assumptions impose any constraints on the observability or any other features of $G[w|\, S=0]$, of the realized treatment distribution in the non-selected subsample.

2.4.1 Monotone Treatment Response

The MTR assumption could also be invoked for the non-selected subsample as it would be reasonable to assume that even children who have dropped out of school at age 15 or lag behind would benefit from a higher maternal education in terms of their test scores, had they stayed in school.

Formally, and as in (14), the MTR assumption states that $\forall\, d, d_1, d_2 \in D$ such that $d_2 > d_1$, $F[y(d_2)|w=d, S=0] \geq_d F[y(d_1)|w=d, S=0]$, that is,

$$F[y(d_2)|w=d, S=0] \leq F[y(d_1)|w=d, S=0]. \tag{23}$$

As was the case with the selected subsample, another channel through which MTR is likely to hold for the non-selected one is that a higher maternal education should weakly increase the probability of school attendance at grade 6 at age 15, that is $E[l(d_2)|w=d, S=0] \geq E[l(d_1)|w=d, S=0]$. This in turn should lead, on average, to weakly higher potential scores $y(d)$. Applying the MTR assumption to both the selected and non-selected subsample implies that the ATE and α-QTE are bounded below by zero. The following lemma establishes the result.

Lemma 1. *Let* $\alpha \in (0,1)$. *Suppose that* $\forall\, d$, and $\forall\, d_1, d_2 \in D$ *such that* $d_2 > d_1$, *and for* $k = 0,1$,

$$F[y(d_2)|w=d, S=k] \leq F[y(d_1)|w=d, S=k]. \tag{24}$$

Then the $ATE(d_2, d_1)$ *and the* α*-*$QTE(d_2, d_1)$ *are bounded below by zero.*

Proof. See Appendix A.2.

As is the case with MTR in the selected subsample, MTR in the non-selected subsample is an untestable assumption. Moreover, since no outcomes are observed in the non-selected subsample, one cannot combine the MTR with the MTS assumption to derive a testable implication using observed outcomes, as is the case with the selected subsample.

2.4.2 Stochastic Dominance

The assumption of stochastic dominance (SD henceforth) compares the distribution of scores under sample selection and no selection and posits that the former distribution stochastically dominates the latter, for any given level of the treatment.

Formally, $\forall\, d \in D$, $F[y(d)|\, S=1] \geq_d F[y(d)|\, S=0]$, that is,

$$F[y(d)|\, S=1] \leq F[y(d)|\, S=0]. \tag{25}$$

In our empirical application, SD states that the distribution of the scores of children who are in school stochastically weakly dominates that of children who have actually dropped out or lag behind, had the latter attended school at grade 6 or above, and had all children had the same level of maternal education. This is a mild assumption, as children that drop out or lag behind are more likely to come from less privileged backgrounds, and thus are likely to do worse in school on average.[3] SD could also obtain if students are more likely to stay in school when they feel that can do well in class, which is another reason for positive selection.

Importantly, SD is likely to hold even if one takes into account the effect of the counterfactual treatment variation on the probability of counterfactual school attendance at grade 6 or above. Especially for higher maternal education levels, such attendance is likely to increase more for observed non-selected children, and this in turn should make their score distribution less stochastically dominated by the one of the selected children. However, it should remain the case that the less privileged socio-economic background observed on average of the non-selected children makes the SD condition in (25) hold.

The SD assumption in (25) implies that a lower bound on $F[y(d)|\, S=1]$ is also a lower bound on $F[y(d)|\, S=0]$. Hence, (6) implies that a lower bound on $F[y(d)|\, S=1]$ is also a lower bound on $F[y(d)]$, as in Blundell et al. (2007). Given that lower bounds on distributions determine

[3] In Brazil, as in most Latin American countries, there is a strong negative association between socio-economic status and secondary school drop-out (Bassi et al., 2015; Busso et al., 2017).

upper bounds on quantiles, SD also implies that the upper bound on $Q_\alpha[y(d)|S=1]$ is also an upper bound on $Q_\alpha[y(d)]$. The same result also holds at the mean, that is, the upper bound on $E[y(d)|\,S=1]$ becomes under SD an upper bound on $E[y(d)]$.

The intuition for these results in our context is that an upper bound on the potential score (under treatment value d) of children who stay in school is also an upper bound for the counterfactual potential score of the children who drop out or lag behind, given that the latter children would have been expected to do worse than the former on average, had they taken the test. As a result, an upper bound on the potential score of children that take the test is also an upper bound on the score of all students.

The above implies that under SD uncertainty due to sample selection no longer affects the upper bounds of $E[y(d)]$ and $Q_\alpha[y(d)]$; as a result, they become smaller. As discussed in Appendix A.4, SD decreases upper bounds the most at higher quantiles of $y(d)$.

2.4.3 Bounded Variation

While SD imposes the condition that non-selected children cannot do better than selected ones in distribution, it does not put any limits on how much worse they can do. Hence, without any further assumptions the conservative lower bound on test scores of non-selected children is y^{INFN}, that is, zero. This is likely an unduly dismal outcome, as at least some non-selected children would have done well in the PISA tests, had circumstances allowed them to continue attending while not lagging behind.

We thus propose an assumption of bounded variation (BV) that limits the extent to which non-selected students would be doing worse in school than the selected ones.[4] This BV assumption, a variant of the BV assumptions used in Manski and Pepper (2018), posits that the lower bound of the mean potential score of the non-selected students cannot be smaller than *k* percent of the lower bound of the mean potential score of students that take the test. In other words, we assume that for the lower bounds SLB_E^L and NLB_E^M of $E[y(d)]$ in (2) we have

$$NLB_E^M = kSLB_E^L. \tag{26}$$

[4] Alternatively, one can view this assumption as one of maximum stochastic dominance, as it limits the extent to which the outcomes in the non-selected subsample are smaller than those in the selected one.

Clearly, it is difficult to pin down exactly the coefficient *k*, and thus we provide results for two values equal to 25 percent and 50 percent in our empirical application (the former is obviously a more conservative choice than the latter). As is also the case with the other identifying assumptions used, we leave it to the reader to decide whether any variant of the BVk assumption is credible.

Turning now to quantiles, we construct the upper bound of the potential outcome distribution of the non-selected children (which determines the lower bound of the associated quantiles) as follows: i) we take the upper bound of the corresponding distribution of selected children and calculate its quantiles from 1 to 99 ; ii) we then make each quantile between 1 and 99 of the upper bound of the potential score distribution in the non-selected subsample *k* percent smaller than the corresponding quantile calculated in i), and then linearly interpolate between the thus calculated quantiles to construct the upper bound of the distribution in the non-selected subsample.

We provide further details on the calculation of bounds under BVk in Appendix A.5. We also point out therein that BVk increases lower bounds the most at lower quantiles of $y(d)$.

2.5 Monotone Instrumental Variables

As MP show, one can further narrow identification regions by using monotone IVs (MIVs henceforth), that is, auxiliary variables allowed to vary weakly monotonically with the outcome.

Formally, a variable Z is a MIV if $\forall\ d \in D, \forall\, z_1, z_2 \in Z$ such that $z_2 > z_1$, $F[y(d)|\, Z = z_2] \geq_d F[y(d)|\, Z = z_1]$, that is,

$$F[y(d)|\, Z = z_2] \leq F[y(d)|\, Z = z_1]. \tag{2722}$$

In other words, the distribution of the potential outcome conditional on a higher value of the MIV must dominate stochastically the corresponding distribution conditional on a lower value of the MIV (see also Blundell et al., 2007: 332).

To understand better how MIVs operate, we first note that we can always express the lower bound on $E[y(d)]$[5] under a set of assumptions L on the selected subsample and a set of assumptions *M* on the non-selected one as

[5] The intuition is the same for the case of the $Q_\alpha[y(d)]$, which we discuss in Appendix A.6.

$$LB_E^{L,M}(d) = \sum_z LB_E^{L,M}(d|Z=z)\, P(Z=z). \tag{28}$$

Clearly, $P(Z=z)$ is given by the data and thus cannot be changed. Hence, to increase the overall lower bound $LB_E^{L,M}(d)$ one needs to increase the conditional lower bounds $LB_E^{L,M}(d|Z=z)$. Similar arguments hold for the upper bound $UB_E^{L,M}(d)$.

Let us first examine how an exogenous IV (XIV) can help narrow the identification range. Following Manski (1990), a variable Z is a XIV if $\forall\, d \in D, \forall\, z \in Z$,

$$E[y(d)|\, Z=z] = E[y(d)]. \tag{29}$$

Equation (29) implies that conditioning on any value of the XIV does not change the distribution of the potential outcome. Hence, all identification regions conditional on values of Z should provide identical lower and upper bounds on $E[y(d)]$. Therefore, the identification region of $E[y(d)]$ is the intersection of all identification regions conditional on Z. This intersection is contained between the maximum of all the conditional lower bounds and the minimum of all conditional upper bounds. Hence, we have

$$\max_z LB_E^{L,M}(d|Z=z) \le E[y(d)] \le \min_z UB_E^{L,M}(d|Z=z). \tag{30}$$

With an MIV, however, equation (29) does not hold because by (27) the MIV is weakly monotonically correlated with the outcome. As a result, one cannot compute as the overall identification region the intersection of all conditional identification regions, as with an XIV. On the other hand, it is possible to exploit the fact that, by (27), a lower bound on $E[y(d)|\, Z=z_1]$ is also a lower bound on $E[y(d)|\, Z=z]$ for $z \ge z_1$, and, correspondingly, an upper bound on $E[y(d)|\, Z=z_2]$ is also a upper bound on $E[y(d)|\, Z=z]$ for $z \le z_2$. Hence, one can potentially increase the lower bound $LB_E^{L,M}(d|Z=z)$ in (28) by taking the maximum lower bound $LB_E^{L,M}(d|Z=z_1)$ over all $z_1 \le z$. Correspondingly, one can potentially decrease the upper bound $UB_E^{L,M}(d|Z=z)$ by taking the minimum upper bound $UB_E^{L,M}(d|Z=z_2)$ over all $z_2 \ge z$. Hence, we obtain

$$\max_{z_1 \le z} LB_E^{L,M}[d|Z=z_1] \le E[y(d)|Z=z] \le \min_{z \le z_2} UB_E^{L,M}[d|Z=z_2] \tag{31}$$

Once the bounds in (31) have been computed for all z, one can take their weighted average over all z and bound the potential outcome $E\big(Y(d)\big)$ as follows:

$$\begin{aligned} &\sum_{z} P(Z=z) \max_{z_1 \le z} LB_E^{L,M}[d|Z=z_1] \\ \le &\sum_{z} P(Z=z) E[y(d)|Z=z] = E[y(d)] \le \\ &\sum_{z} P(Z=z) \min_{z \le z_2} UB_E^{L,M}[d|Z=z_2] \end{aligned} \tag{32}$$

Hence, by integrating Z out of the conditional expectation $E[y(d)|Z=z]$ one can obtain bounds on $E[y(d)]$. Clearly, the optimization operations in (31) take place over a restricted range of values of Z compared to (29), and thus the identifying power of the MIV assumption is smaller than that of the XIV one. This is to be expected, as the weak monotonicity of a MIV in (27) is a weaker assumption than the exogeneity of an XIV in (29).

We note that (27) does not imply a causal effect of the MIV on the outcome, but only denotes a weakly positive association with it. Importantly, no association is also allowed. We also note that, the MIV condition in (27) is applied to the distribution of the potential outcome in the population, and not just in the selected subsample. As it is applied to potential outcomes, it is an unverifiable assumption, which is also the case for the exogeneity assumption underlying XIVs.

It is generally easier to find candidate MIVs, as the weak stochastic dominance condition in (27) is less demanding than the orthogonality with the outcome that is required of an XIV. In our application, we use two monotone instruments. The first one is the father's education, measured in three levels, primary, secondary and tertiary. The MIV condition implies that we require that a higher paternal education is not negatively associated with school performance in the population. This seems a mild assumption in our case, and its plausibility is further discussed in Section 3.

The second MIV we use is an indicator of material possessions, and in particular of a family car and a computer at home. This indicator takes the value of zero if none of the goods are owned by the child's family, one if the family possesses any one of the two goods, and two if it possesses both. The MIV assumption requires that possessing more of these two goods is not negatively associated with school performance in the population. Once more, this is a mild assumption given that these two goods denote socio-economic status. We discuss this further in Section 3.

3. Empirical Analysis: The Impact of Maternal Education on School Achievement

3.1 Data

In our empirical application, we use the mathematics test scores reported in the 2015 PISA survey in Brazil. Our sample consists of 17,984 children aged 15 who are administered the test at school. The PISA survey also reports the education of the mother, which we divide in three levels: primary, secondary and tertiary. In our application we report the treatment effect of a change in maternal education from primary to tertiary. The average test score is equal to 385.8 points, while the median is 380.5 points, and the standard deviation 91.3 points.

We further observe that there is a clear positive association between the test score and maternal education, both at the mean and the median. In particular, the mean scores for schoolchildren whose mothers have primary, secondary and tertiary education are 361, 393 and 421 points, respectively. The corresponding results at the median are equal to 359, 391 and 421 points, respectively. We obtain analogous results in other quantiles. This strong positive association between the observed outcome and the treatment implies that it is not possible to reject the joint MTR+MTS hypothesis in the selected subsample, as discussed in Section 2.3.2 above.

The PISA data do not contain any information on the percentage of students who have dropped out of school or are in grade 6 or below at age 15, and hence are excluded from the PISA sampling frame. To remedy this situation, we obtain data on school enrolment from the Pesquisa Nacional por Amostra de Domicílios (PNAD), a nationally representative household survey. The PNAD contains information on whether children are currently attending school and the grade attended, which allows us to construct a selection indicator adhering to the PISA sample selection criteria.

The PNAD also includes information on the treatment (maternal education), and, importantly, on our two MIVs, namely paternal education and family ownership of a computer and a car.[6] As discussed above, we do not use the information on the distribution of the treatment in the non-selected subsample $G[w|\, S = 0]$ in our estimation, as none of our identifying assumptions is related to it. On the other hand, we use the information on the two MIVs, as we need to compute the probability of selection conditional on them. We pool three years of PNAD

[6] Parental education is not available for 15-year-olds who are heads of household or their spouses in the survey (2.8 percent of the 15-year-olds included in the three survey years—2013, 2014, 2015—used for the analysis).

data, that is, from 2013 to 2015, to obtain more precise estimates of these selection probabilities. The final sample consists of 9,712 children aged 15.

Using PNAD, we compute the probability of selection and its complement. Some 13.9 percent of eligible 15-year-olds are not included in PISA, and this selection probability varies strongly by the level of maternal education: it is equal to 21 percent for children who have mothers with primary education, 7.7 percent for those with mothers with secondary education, and 4 percent for those with mothers with tertiary education. The strong positive association between maternal education and PISA selection makes it more likely that the assumption of stochastic dominance discussed in Section 2.4.2 above is a valid one.

3.2 Statistical Considerations

As Kreider and Pepper (2007) and Manski and Pepper (2009) point out, the minimization and maximization operations required by the MIV assumption can result in biased estimates. To correct for this, we use for our mainline results the bias correction procedure proposed by Kreider and Pepper (2007). The bias-corrected estimates turned out to be very close to the uncorrected ones (the latter are available upon request from the authors).

For methods that do not use MIVs we compute confidence intervals (CIs henceforth) using the results in Imbens and Manski (2004). On the other hand, for methods that do use MIVs, Manski and Pepper (2009: S211) point out that the methods underlying the Imbens and Manski (2004) CIs appear to be inapplicable. Therefore, we use, as in de Haan (2011) bootstrap-based bias-corrected percentile CIs. We use 80 bootstrap replications, and the associated balanced replicate weights, as recommended in OECD (2017b, Chapter 8).

In our mainline results, we do not bootstrap the selection rates derived from the PNAD data. As a robustness check, however, we bootstrap those rates as well by resampling the PNAD sample 80 times (so as to match the number of bootstrap runs in the PISA data). The resulting CIs remain essentially unchanged.

3.3 Results

We show results on the treatment effects on test scores due to a change in maternal education from primary to tertiary. In each set of results, we show the lower and upper bound of the treatment effect, as well as the associated 95 percent and 90 percent CIs. To highlight the role of the

uncertainty induced by sample selection, we juxtapose the estimates of the population ATEs and α-QTEs under sample selection to those for the ATEs and α-QTEs in the selected subsample, which consists of the students taking the PISA test. The latter estimates are derived using results in Manski (1997), MP, and Giustinelli (2011).

Panel A of Table 2 shows the estimates of the average and median treatment effects in the population, while the corresponding results for the selected subsample are shown in Panel B. Each row shows results obtained under a particular set of assumptions. Column 1 shows the assumptions used for the computation of the bounds in the selected sample, while Column 2 shows the assumptions used in the non-selected sample. Column 3 indicates whether MIVs are used. Treatment effects at the mean are shown in Columns 4-9, while those at the median in Columns 10-15.

We start with the results for treatment effects computed under the assumption that sample selection is random and the treatment is exogenous (ETS). If both these assumptions hold, then the causal impact of maternal education is given by the difference in observed average and median scores in the PISA data. We note that under the ETS assumption maternal education has a strong influence on children's school performance, namely 59.2 points at the mean and 62.7 points at the median, with both effects being tightly estimated.

The ETS results are likely misleading, however, both because maternal education is unlikely to be exogenous and because sample selection (i.e., dropping out of school and lagging behind) is unlikely to be random. As discussed in Section 2.1, to address these two problems we use PI methods. It is clear that under NA, that is, the most conservative estimation choice, the identification region of the treatment effect is very wide and thus uninformative, as it ranges from -459.6 to 501.8 points at the mean and from -741.9 to 741.9 points at the median. The MTR assumption applied to the selected subsample raises considerably the lower bound of the treatment effect but leaves it still well below zero, while adding the MTR assumption also on the non-selected subsample raises further the lower bound to zero, as shown in Lemma 1 in Section 2.4.1. In both cases the upper bound remains unchanged.

Turning now to the MTS assumption, when used on its own on the selected subsample and together with NA in the non-selected one it produces an uninformative lower bound. On the other hand, the upper bound shrinks considerably compared to the one under NA in both subsamples,

namely to 154.7 points at the mean, and 101.3 points at the median. Adding the MTR assumption in both subsamples to the MTS one brings back the lower bounds to zero.

Combining MTS and MTR for the selected subsample with MTR and SD for the non-selected one narrows down the identification range considerably, with the lower bounds being equal to zero while the upper bounds shrink to 110 points at the mean and 78.2 points at the median.

Adding the BV25 assumption to the previous combination decreases considerably the upper bound of the ATE to 97.5 points, while the BV50 assumption decreases it further to 84.9 points. On the other hand, as mentioned in Section 2.6.3 and further explained in Appendix A.5 the BV assumption narrows very little the identification regions at quantiles other than the very low ones, and thus adding even the stronger BV50 assumption decreases the upper bound of the median QTE by only 0.3 points with respect to that obtained under MTR+SD.

Adding the MIV1 assumption (i.e., only using paternal education as a MIV) to the MTR and MTS assumptions for the selected subsample, and the MTR and SD assumptions for the non-selected one, makes the upper bounds shrink to 85.9 points at the mean and 55.1 points at the median. Adding the BV25 assumption shrinks the upper bound at the mean to 74.7 points, while the BV50 assumption shrinks it further to 63.6 points. On the other hand, the BVk assumptions make again little difference at the median. In all cases, the lower bound of the treatment effect is equal to zero.

Adding a second monotone instrument (i.e. the family ownership of a car or a computer) shrinks the identification region further, as the upper bound when using MTR and MTS for the selected subsample and MTR and SD for the non-selected one becomes equal to 63.5 points at the mean and 30.6 points at the median, that is, equal to about 0.34 standard deviations (sds) of the score. Once more, adding the BVk assumptions makes little difference at the median, while BV25 shrinks the upper bound at the mean to 53.5 points, and BV50 to 43.4 points, which is about 0.48 sds of the observed score distribution. Once more, the lower bounds are equal to zero.

The above results clearly show that the identification region of the ATEs and QTEs narrows as assumptions are added, and the identifying power of each additional assumption becomes clear when comparing results with and without it. The smallest upper bounds we obtain at the mean are considerably lower than the ETS estimates, while at the median this happens when adding the MIV1 assumption to the MTR and MTS ones for the selected subsample, and the SD one for the non-selected one. The fact that the ETS results are well outside the PI identification

regions obtained using mild assumptions suggests that they are likely to overestimate the causal effect of maternal education on the child's math score. Importantly, the uncertainty around the PI estimates is also much larger than the one around the ETS ones, suggesting that ETS and the random sample selection assumption lead to a considerable underestimation of the true uncertainty around the effect of interest.

In Panel B of Table 2 we show the treatment effects estimated using only the selected subsample (the test-takers). Hence, these estimates do not reflect what happens in the population, but only in this subsample. Given that the uncertainty due to the unobservability of test scores of non-test-takers is ignored, we expect the identification regions in this subsample to be narrower.

Indeed, we observe that the same assumptions in the selected subsample lead to much narrower identification regions when ignoring sample selection. For example, under MTR and MTS, the upper bounds are equal to 59.8 points at the mean, and 62.7 points at the median.[7] The corresponding upper bounds of population treatment effects, even after adding MTR and SD on the non-selected subsample, are much larger at 110 points, and 78.2 points, as noted above.

When we add the MIV assumption the identification regions shrink even further, with the lower bounds becoming equal to 1.2 points at the mean and 2.7 points at the median when using MTR+MTS+MIV2. Both lower bounds are statistically significant at 5%, and imply that changing maternal education from primary to tertiary causes an increase in the math score that is at least this large. The upper bounds also become much smaller than those computed without the MIV assumption: they are equal to 23.1 points at the mean and 20.7 points at the median, that is, about 0.25 and 0.23 sds of the observed score distribution, respectively. Both these values are much smaller than the ETS estimates, which again suggests that ETS overestimates the effect of maternal education on the child's score also in the selected subsample, that is, for the children who actually take the test.

In Table 3 we show results for the 10^{th} and the 90^{th} quantiles. We note that also for these quantiles the addition of assumptions generally narrows the identification regions. At the 10^{th} quantile (results are shown in Columns 4-9), however, the upper bounds remain very large. This is due to the fact that, as shown in the definition of the α-QTE given in (7), the upper bound of the

[7] These values are equal to the ETS ones. The result that the MTR+MTS upper bound of the treatment effect of a change in the treatment from its minimum to its maximum value is equal to the ETS value is shown by MP for the ATE, and by Giustinelli (2011) for the QTE.

treatment effect is equal to the upper bound of the 10th quantile of the potential outcome distribution under tertiary maternal education minus the lower bound of the 10th quantile of the potential outcome distribution under primary maternal education. It turns out that the bound in the latter case is the smallest possible, that is zero, and thus the upper bound of the 10-QTE becomes very large. The lower bound at the 10th quantile of the potential outcome distribution under primary maternal education is so small due to the fact that, as discussed in Section 3, in our sample the probability of non-selection is larger than 0.1, and thus Propositions 1-3 imply that the lower bound at the 10th quantile is not identified. Consequently, it is equal to min(y^{INFS}, y^{INFN}), that is, zero.

Only when applying the MIV and BVk assumptions (see the discussion in Section 2.4.3 and in Appendix A.5 for the identifying power of the latter at low quantiles) does the upper bound of the 10-QTE narrow considerably, but even at its lowest (equal to 69.5 points) it remains much larger than the ETS estimate, which is equal to 24.5 points. The lower bound of the 10-QTE is always zero.

The 10-QTEs on the selected sample are shown in Panel B, and, as expected, have identification regions that are much narrower than when incorporating sample selection. The upper bound under MTR+MTS is equal to 24.5, that is to the ETS value, while adding two MIVs makes the lower bound equal to 2.6 (significant at 5 percent), while the upper bound becomes 4.3. Hence, the MIV assumption results in a very narrow identification whose upper bound is much smaller the ETS value, suggesting again that the latter is an upwardly biased estimate of the 10-QTE for the test takers.

Turning now to the 90-QTE (results are shown in Columns 10-15 in Table 3), we see that the SD assumption, as discussed in Section 2.4.2 and Appendix A.4, has considerable identifying power at high quantiles. On the other hand, the BVk assumption has very little identifying power, as discussed in Section 2.4.3 and in Appendix A.5. Without MIV, the narrowest upper bound is equal to 97.4 points, while adding MIV2 (MIV1) brings it down to 45.4 (45.2) points, that is, about 0.5 sds of the observed score distribution. Hence, it is much smaller than the ETS value (89.3 points). The lower bounds of the 90-QTE are always equal to zero.

In the selected subsample (results are shown in Panel B of Table 3), the narrowest upper bound using MTR+MTS+MIV2 is 43.2 points, while the corresponding one under

MTR+MTS+MIV1 is 61.1 points. Hence, at its tightest, the upper bound is equal to about 0.47 sds of the observed score distribution, and well below the ETS value.

In Appendix Table A.1 we provide results for the 25-QTE and the 75-QTE. The patterns of the results are broadly similar to those for the quantiles discussed up to now. One important difference of the 25-QTE with respect to the 10-QTE is that the upper bounds of the former are now much smaller than those of the latter. This happens because the probability of non-selection in our sample is smaller than 0.25, and thus the lower bound of the 25th quantile of the potential outcome distribution under primary maternal education is identified, as indicated by Propositions 1-3.

4. Discussion

The need to estimate treatment effects in the population when the available sample is a selected one is a common occurrence in empirical work. Ignoring sample selection likely leads to biased estimates of population treatment effects and underestimates their standard errors, given the additional uncertainty induced by the unobservability of outcomes in parts of the population.

In this paper, we address these issues by applying PI methods that use mild assumptions to bound nonparametrically population ATEs and QTEs under sample selection. In the process, we show the extent to which each additional assumption narrows identification regions. In contrast with previous literature that has applied similar methods to situations where treatment assignment is random (i.e., in the context of randomized control trials), we derive these regions in observational data, where this is generally not the case.

We apply our procedures to the estimation of the causal effect of maternal education on the child's math score in the 2015 PISA test in Brazil. The test is administered to 15-year-old children attending secondary school at grade 6 or above. Using data from PNAD, a nationally representative sample of Brazilian households, we show that administering the test in schools misses an important part of the target population: 13.9 percent of 15-year-old children have dropped out from school or attend a grade below 6. Sample selection is non-random: the non-selection rate is 21 percent for children who have mothers with primary education, and 4 percent for those with mothers with tertiary education.

Ignoring this non-random sample selection, that is, examining only the sample of test takers, leads to narrow identification regions and, depending on the set of assumptions used,

treatment effects of maternal education on children's math test scores that are significantly different from zero. The narrowest identification regions imply that having a mother with tertiary education increases the score by at least about 0.01(0.03) standard deviations at the mean(median) compared to having a mother with primary education. The upper bound of the effect is equal to about 0.25(0.23) standard deviations at the mean (median).

On the other hand, when taking into account the additional uncertainty due to sample selection, even the narrowest identification regions we obtain imply that one cannot reject the hypothesis that the population treatment effect is zero. In contrast, the effect can be considerable at its upper bound, equal to about 0.48(0.36) standard deviations of the observed score at the mean(median). Moreover, identification regions are much larger at the bottom of the score distribution.

Importantly, the upper bounds of treatment effects are always lower than the observed difference in scores among children with different levels of maternal education (with the exception of the 10-QTE). This is to be expected, as interpreting observed differences causally implies treating maternal education as exogenous to the test score and sample selection (i.e., test-taking) as random.

The identification regions could become narrower if there were some additional information on the non-selected subsample. For example, if tests like PISA were administered to at least some of the non-selected children, then researchers would have a better idea of the difference in scores between those children and the children that still attend school at grade 6 or higher at age 15. We understand that such surveys on school dropouts are currently being considered, and their eventual availability could reduce the uncertainty about population treatment effects of interest.

References

Ashenfelter, O., and J. Ham. 1979. “Education, Unemployment, and Earnings.” *Journal of Political Economy* 87: 99-116.

Bassi, M., M. Busso and J.S. Muñoz. 2015. “Enrollment, Graduation, and Dropout Rates in Latin America: Is the Glass Half Empty or Half Full?” *Economía* 16(1): 113-156.

Blanco, G., C. Flores and A. Flores-Lagunes. 2012. “Bounds on Average and Quantile Treatment Effects of Job Corps Training on Wages.” *Journal of Human Resources* 48: 659-701.

Blundell, R. et al. 2007. “Changes in the Distribution of Male and Female Wages Accounting for Employment Composition Using Bounds.” *Econometrica* 75: 323–363.

Busso, M. et al. 2017. *Learning Better: Public Policy for Skills Development.* Development in the Americas report. Washington, DC: United States: Inter-American Development Bank.

Chen, X., C.A. Flores and and A. Flores-Lagunes. 2018. “Going beyond LATE: Bounding Average Treatment Effects of Job Corps Training.” *Journal of Human Resources* 53(4): 1050-1099.

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

De Haan, M. 2011. “The Effect of Parents’ Schooling on Child’s Schooling: A Nonparametric Bounds Analysis.” *Journal of Labor Economics* 29: 859–892.

Giustinelli, P. 2011. “Non-Parametric Bounds on Quantiles under Monotonicity Assumptions: With an Application to the Italian Education Returns.” *Journal of Applied Econometrics* 26: 783–824.

Gronau, R. 1973. “The Effect of Children on the Housewife’s Value of Time.” *Journal of Political Economy* 81: S168–S199.

Heckman, J.J. 1974. “Shadow Prices, Market Wages and Labor Supply.” *Econometrica* 42: 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* 5: 475–492.

Heckman, J.J. 1979. “Sample Selection Bias as a Specification Error.” *Econometrica* 47: 153–161.

Holmlund, H., M. Lindahl, M., and E. Plug. 2011. “The Causal Effect of Parents’ Schooling on Children’s Schooling: A Comparison of Estimation Methods.” *Journal of Economic Literature* 49(3): 615–651.

Horowitz, J.L., and C.F. Manski. 2000. “Nonparametric Analysis of Randomized Experiments with Missing Covariate and Outcome Data.” *Journal of the American Statistical Association* 95: 77–84.

Huber, M., and G. Mellace. 2015. “Sharp Bounds on Causal Effects under Sample Selection.” *Oxford Bulletin of Economics and Statistics* 77: 129–151.

Huber, M., L. Laffers and G. Mellace. 2017. “Sharp IV Bounds on Average Treatment Effects on the Treated and Other Populations under Endogeneity and Noncompliance.” *Journal of Applied Econometrics* 32: 56-79.

Imbens, G.W., and C. Manski. 2004. “Confidence Intervals for Partially Identified Parameters.” *Econometrica* 72: 1845–57.

Kreider, B., and J.V. Pepper. 2007. “Disability and Employment: Re-evaluating the Evidence in Light of Reporting Errors.” *Journal of the American Statistical Association* 102: 432–41.

Lee, L.F. 1982. “Some Approaches to the Correction of Selectivity Bias.” *Review of Economic Studies* 49: 355–372.

Lee, D. 2009. “Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects.” *Review of Economic Studies* 76(3): 1071–1102.

Manski, C.F. 1989. “Anatomy of the Selection Problem.” *Journal of Human Resources* 24: 343–60.

Manski, C.F. 1990. “Nonparametric Bounds on Treatment Effects.” *American Economic Review* 80: 319–323.

Manski, C.F. 1994. “The Selection Problem.” In: C. Sims, editor. *Advances in Econometrics: Sixth World Congress*. Cambridge, United Kingdom: Cambridge University Press.

Manski, C.F. 1997. “Monotone Treatment Response.” *Econometrica* 65: 1311–1334.

Manski, C.F., and J. Pepper. 2000. “Monotone Instrumental Variables: With an Application to the Returns to Schooling.” *Econometrica* 68: 997–1010.

Manski, C.F., and J. Pepper. 2009. “More on Monotone Instrumental Variables.” *Econometrics Journal* 12: S200–S216.

Manski, C.F., and J. Pepper. 2018. “How Do Right-to-Carry Laws Affect Crime Rates? Coping with Ambiguity Using Bounded Variation Assumptions.” *Review of Economics and Statistics* 100: 232–244.

Molinari, F. 2012. “Missing Treatments.” *Journal of Business and Economics Statistics* 28: 82–95.

Newey, W.K. 2009. “Two-Step Series Estimation of Sample Selection Models.” *Econometrics Journal* 12: S217-229.

Organisation for Economic Co-operation and Development (OECD). 2017a. *PISA 2015 Assessment and Analytical Framework: Science, Reading, Mathematic, Financial Literacy and Collaborative Problem Solving*. Paris, France: OECD Publishing.

----. 2017b. *PISA 2015 Technical Report*. Paris, France: OECD Publishing.

Table 1. Descriptive Statistics

Magnitude	Mean	Median	Std. Deviation	Minimum	Maximum
PISA math test Score					
Whole sample	385.79	380.47	91.26	88.34	741.90
Maternal education less than secondary	361.29	358.53	80.33	100.07	671.40
Maternal education secondary	393.41	391.49	87.20	88.34	739.55
Maternal education tertiary	421.06	421.27	102.43	111.46	741.90
Number of observations in PISA	17,984				
School Enrolment					
Whole Sample	0.861				
Maternal education less than secondary	0.790				
Maternal education secondary	0.923				
Maternal education tertiary	0.960				
Number of observations in PNAD	9,712				

Table 2. Average and Median Treatment Effects

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
Assumptions			Mean						Median					
Assumptions used for the selected subsample	Assumptions used for the non-selected subsample	MIV	Lower Bound	Upper Bound	Low 95% CI	High 95% CI	Low 90% CI	High 90% CI	Lower Bound	Upper Bound	Low 95% CI	High 95% CI	Low 90% CI	High 90% CI
Panel A. Population treatment effects														
ETS	Random sample selection	No	59.8		53.9	63.3	54.8	62.7	62.7		56.8	67.6	58.3	66.0
NA	NA	No	-459.6	501.8	-463.8	504.5	-462.9	503.9	-741.9	741.9	-741.9	741.9	-741.9	741.9
MTR	NA	No	-103.2	501.8	-103.2	504.5	-103.2	503.9	-37.6	741.9	-39.8	741.9	-39.3	741.9
MTR	MTR	No	0.0	501.8	0.0	504.5	0.0	503.9	0.0	741.9	0.0	741.9	0.0	741.9
MTS	NA	No	-459.6	154.7	-463.8	161.2	-462.9	159.8	-741.9	101.3	-741.9	110.6	-741.9	108.6
MTR+MTS	MTR	No	0.0	154.7	0.0	161.2	0.0	159.8	0.0	101.3	0.0	110.6	0.0	108.6
MTR+MTS	MTR+SD	No	0.0	110.0	0.0	117.6	0.0	115.9	0.0	78.2	0.0	86.9	0.0	85.0
MTR+MTS	MTR+SD+BV25	No	0.0	97.5	0.0	105.0	0.0	103.4	0.0	78.0	0.0	86.7	0.0	84.7
MTR+MTS	MTR+SD+BV50	No	0.0	84.9	0.0	92.5	0.0	90.8	0.0	77.9	0.0	86.6	0.0	84.7
MTR+MTS	MTR+SD	MIV1	0.0	85.9	0.0	89.0	0.0	88.4	0.0	55.1	0.0	59.1	0.0	58.6
MTR+MTS	MTR+SD+BV25	MIV1	0.0	74.7	0.0	78.0	0.0	77.3	0.0	54.9	0.0	58.9	0.0	58.4
MTR+MTS	MTR+SD+BV50	MIV1	0.0	63.6	0.0	66.9	0.0	66.2	0.0	54.8	0.0	58.8	0.0	58.4
MTR+MTS	MTR+SD	MIV2	0.0	63.5	0.0	67.8	0.0	66.9	0.0	30.8	0.0	35.7	0.0	35.7
MTR+MTS	MTR+SD+BV25	MIV2	0.0	53.5	0.0	57.8	0.0	56.8	0.0	30.6	0.0	35.5	0.0	35.5
MTR+MTS	MTR+SD+BV50	MIV2	0.0	43.4	0.0	47.9	0.0	46.8	0.0	30.6	0.0	35.5	0.0	35.5
Panel B. Treatment effects for the selected subsample														
NA	-..-	No	-414.0	463.0	-418.9	466.2	-417.8	465.5	-655.9	655.9	-655.9	655.9	-655.9	655.9
MTR	-..-	No	0.0	463.0	0.0	466.2	0.0	465.5	0.0	655.9	0.0	655.9	0.0	655.9
MTS	-..-	No	-414.0	59.8	-418.9	67.4	-417.8	65.7	-655.9	62.7	-655.9	71.4	-655.9	69.5
MTR+MTS	-..-	No	0.0	59.8	0.0	67.4	0.0	65.7	0.0	62.7	0.0	71.4	0.0	69.5
MTR+MTS	-..-	MIV1	0.0	41.3	0.0	44.7	0.0	44.1	0.0	42.2	0.0	46.4	0.0	44.9
MTR+MTS	-..-	MIV2	1.2	23.1	0.5	27.7	0.6	26.5	2.6	20.7	1.5	25.3	1.6	25.3

Note: The table shows the treatment effect of a change in maternal education from primary to tertiary under different assumptions. ETS: exogenous treatment selection; NA: no assumptions; MTR: monotone treatment response; MTS: monotone treatment selection; SD: stochastic dominance; BV25: variation bounded to 25 percent; BV50: variation bounded to 50 percent: MIV1: single monotone instrumental variable (paternal education); MIV2: two monotone instrumental variables (paternal education and assets ownership, namely having a car and a computer at home). Data sources: PISA and PNAD.

Table 3. Treatment Effects: 10th and 90th Quantiles

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
Assumptions			10th quantile						90th quantile					
Assumptions used for the selected subsample	Assumptions used for the non-selected subsample	MIV	Lower Bound	Upper Bound	Low 95% CI	High 95% CI	Low 90% CI	High 90% CI	Lower Bound	Upper Bound	Low 95% CI	High 95% CI	Low 90% CI	High 90% CI
Panel A. Population treatment effects														
ETS	Random sample selection	**No**	24.6		20.5	28.2	21.7	28.0	89.0		76.5	93.7	79.0	91.8
NA	**NA**	**No**	-308.7	422.9	-313.7	431.4	-312.6	429.5	-323.2	332.4	-339.7	336.9	-336.1	335.9
MTR	**NA**	**No**	-279.6	422.9	-283.8	431.4	-282.9	429.5	-244.3	332.4	-253.1	336.9	-251.2	335.9
MTR	**MTR**	**No**	0.0	422.9	0.0	431.4	0.0	429.5	0.0	332.4	0.0	336.9	0.0	335.9
MTS	**NA**	**No**	-308.7	293.9	-313.7	300.4	-312.6	299.0	-323.2	282.8	-339.7	288.9	-336.1	287.6
MTR+MTS	**MTR**	**No**	0.0	293.9	0.0	300.4	0.0	299.0	0.0	282.8	0.0	288.9	0.0	287.6
MTR+MTS	**MTR+SD**	**No**	0.0	286.1	0.0	292.1	0.0	290.8	0.0	97.9	0.0	110.8	0.0	108.0
MTR+MTS	**MTR+SD+BV25**	**No**	0.0	184.9	0.0	190.7	0.0	189.4	0.0	97.5	0.0	110.5	0.0	107.6
MTR+MTS	**MTR+SD+BV50**	**No**	0.0	93.2	0.0	99.1	0.0	97.8	0.0	97.4	0.0	110.4	0.0	107.5
MTR+MTS	**MTR+SD**	**MIV1**	0.0	278.4	0.0	282.7	0.0	282.4	0.0	65.7	0.0	71.6	0.0	70.2
MTR+MTS	**MTR+SD+BV25**	**MIV1**	0.0	170.8	0.0	174.9	0.0	174.4	0.0	65.5	0.0	71.5	0.0	70.1
MTR+MTS	**MTR+SD+BV50**	**MIV1**	0.0	78.1	0.0	82.6	0.0	81.0	0.0	65.5	0.0	71.4	0.0	70.0
MTR+MTS	**MTR+SD**	**MIV2**	0.0	274.2	0.0	275.8	0.0	275.6	0.0	45.8	0.0	51.8	0.0	51.5
MTR+MTS	**MTR+SD+BV25**	**MIV2**	0.0	161.9	0.0	163.7	0.0	163.2	0.0	45.4	0.0	51.3	0.0	50.8
MTR+MTS	**MTR+SD+BV50**	**MIV2**	0.0	69.5	0.0	72.3	0.0	71.6	0.0	45.2	0.0	51.2	0.0	50.3
Panel B. Treatment effects for the selected subsample														
NA	**-..-**	**No**	-214.3	315.7	-219.5	321.5	-218.3	320.2	-301.9	321.4	-315.9	324.9	-312.8	324.1
MTR	**-..-**	**No**	0.0	315.7	0.0	321.5	0.0	320.2	0.0	321.4	0.0	324.9	0.0	324.1
MTS	**-..-**	**No**	-214.3	24.6	-219.5	31.0	-218.3	29.6	-301.9	89.0	-315.9	101.6	-312.8	98.8
MTR+MTS	**-..-**	**No**	0.0	24.6	0.0	31.0	0.0	29.6	0.0	89.0	0.0	101.6	0.0	98.8
MTR+MTS	**-..-**	**MIV1**	0.0	12.9	0.0	17.4	0.0	17.0	0.0	61.1	0.0	66.5	0.0	65.8
MTR+MTS	**-..-**	**MIV2**	2.6	4.3	0.8	6.0	0.9	6.0	0.0	43.2	0.0	49.8	0.0	48.6

Note: See note to Table 2.

Appendix

A.1 No Assumptions (NA) Bounds

Proof of Proposition 1. The proof is based on Manski (1994: 149-151) and is structured in four parts.

1. $s^{NA,NA}(\alpha, d)$ is an upper bound.

The NA,NA bounds on $F[y(d)]$ in (10) imply that

$$F[y(d) \leq \tilde{y}|w = d, S = 1]P(w = d|S = 1)P(S = 1) \geq \alpha \implies F[y(d) \leq \tilde{y}] \geq \alpha. \quad \text{(A.1)}$$

The premise of (A.1) is empty if $P(w = d|S = 1)P(S = 1) < \alpha$. Suppose that $P(w = d|S = 1)P(S = 1) \geq \alpha$. Then the definition of $s^{NA,NA}(\alpha, d)$ states that

$$s^{NA,NA}(\alpha, d) \equiv min\,\tilde{y}: F[y(d) \leq \tilde{y}|w = d, S = 1] \geq \alpha/[P(w = d|S = 1)P(S = 1)]. \quad \text{(A.2)}$$

It follows that $F[y(d) \leq s^{NA,NA}(\alpha, d)] \geq \alpha$. Hence $Q_\alpha[y(d)] \leq s^{NA,NA}(\alpha, d)$.

This result can be understood from the fact that the upper bound on $Q_\alpha[y(d)]$ is found by reversing the NA,NA lower bound on $F[y(d)]$ in (10). Hence, it is the value $\tilde{y}$ that solves the equation $F(\tilde{y}|w = d, S = 1)P(w = d|S = 1)P(S = 1) = \alpha \Longrightarrow F(\tilde{y}|w = d, S = 1) = \frac{\alpha}{P(w=d|S=1)P(S=1)} = \beta$. Hence, $\tilde{y}$ is equal to the β^{th} quantile of $F(y|w = d, S = 1)$, and thus it must be the case that $\beta \leq 1 \Longrightarrow P(w = d|S = 1)P(S = 1) \geq \alpha$.

2. $r^{NA,NA}(\alpha, d)$ is a lower bound.

The NA,NA bounds on $F[y(d)]$ in (10) imply that

$$\{F[y(d) \leq \tilde{y}|w = d, S = 1]P(w = d|S = 1) + P(w \neq d|S = 1)\}P(S = 1) + P(S = 0) < \alpha \Longrightarrow F[y(d) \leq \tilde{y}] < \alpha. \quad \text{(A.3)}$$

Given that $P(w \neq d|S = 1) = 1 - P(w = d|S = 1)$ and $P(S = 0) = 1 - P(S = 1)$, (A.3) may be rewritten as

$$F[y(d) \leq \tilde{y}|w = d, S = 1] < 1 - \frac{(1-\alpha)}{P(w=d|S=1)P(S=1)} \Longrightarrow F[y(d) \leq \tilde{y}] < \alpha. \quad \text{(A.4)}$$

The premise of (A.4) is empty if $1 - \alpha \geq P(w = d|S = 1)P(S = 1) \Longrightarrow P(w \neq d|S = 1)P(S = 1) + P(S = 0) \geq \alpha$. Suppose that $1 - \alpha < P(w = d|S = 1)P(S = 1) \Longrightarrow P(w \neq d|S = 1)P(S = 1) + P(S = 0) < \alpha$. Then the definition of $r^{NA,NA}(\alpha, d)$ states that

$$r^{NA,NA}(\alpha, d) \equiv \min \tilde{y}: F[y(d) \leq \tilde{y}|w = d, S = 1] \geq 1 - \left\{\frac{(1-\alpha)}{P(w = d|S = 1)P(S = 1)}\right\}. \quad \text{(A.5)}$$

It follows that, for all $\eta > 0$, $F[y(d) \leq r^{NA,NA}(\alpha, d) - \eta] < \alpha$. Hence $r^{NA,NA}(\alpha, d) \leq Q_\alpha[y(d)]$.

This result can be understood from the fact that the lower bound on $Q_\alpha[y(d)]$ is found by reversing the NA,NA upper bound on $F[y(d)]$ in (10). Hence, it is the value $\tilde{y}$ that solves the equation $[F(\tilde{y}|w = d, S = 1)P(w = d|S = 1) + P(w \neq d|S = 1)]P(S = 1) + P(S = 0) = \alpha \Longrightarrow F(\tilde{y}|w = d, S = 1) = 1 - \frac{1\text{-}\alpha}{P(w=d|S=1)P(S=1)} = \beta$. Hence, $\tilde{y}$ is equal to the β^{th} quantile of $F(y|w = d, S = 1)$, and thus it must be the case that $\beta > 0 \Longrightarrow 1 - \alpha < P(w = d|S = 1)P(S = 1) \Longrightarrow P(w \neq d|S = 1)P(S = 1) + P(S = 0) < \alpha$.

3. $s^{NA,NA}(\alpha, d)$ is the least upper bound.

Let $P(w = d|S = 1)P(S = 1) \geq \alpha$, so that $s^{NA,NA}(\alpha, d)$ is finite. For any $\lambda > 0$,

$$\begin{aligned} F[y(d) \leq s^{NA,NA}(\alpha, d) - \lambda] = {} & F[y(d) \leq s^{NA,NA}(\alpha, d) - \lambda|S = 1]P(S = 1) + \\ & F[y(d) \leq s^{NA,NA}(\alpha, d) - \lambda|S = 0]P(S = 0) = \\ & \{F[y \leq s^{NA,NA}(\alpha, d) - \lambda|w = d, S = 1]P(w = d|S = 1) + \\ & F[y(d) \leq s^{NA,NA}(\alpha, d) - \lambda|w \neq d, S = 1]P(w \neq d|S = 1)\}P(S = 1) + \\ & F[y(d) \leq s^{NA,NA}(\alpha, d) - \lambda|S = 0]P(S = 0). \end{aligned} \tag{A.6}$$

Suppose that $F[y(d) \leq s^{NA,NA}(\alpha, d) - \lambda|w \neq d, S = 1] = F[y(d) \leq s^{NA,NA}(\alpha, d) - \lambda|S = 0] = 0$, as is possible in the absence of prior information. Then, the definition of $s^{NA,NA}(\alpha, d)$ implies that

$$\begin{aligned} & F[y(d) \leq s^{NA,NA}(\alpha, d) - \lambda] = \\ & F[y \leq s^{NA,NA}(\alpha, d) - \lambda|w = d, S = 1]P(w = d|S = 1)P(S = 1) < \alpha. \end{aligned} \tag{A.7}$$

Hence, $Q_\alpha[y(d)] > s^{NA,NA}(\alpha, d) - \lambda$.

Suppose now that $P(w = d|S = 1)P(S = 1) < \alpha$, so that $s^{NA,NA}(\alpha, d) = max(\, y^{SUPS}, \, y^{SUPN})$. For any finite t,

$$\begin{aligned} F[y(d) \leq t] = {} & F[y(d) \leq t|S = 1]P(S = 1) + F[y(d) \leq t|S = 0]P(S = 0) = \\ & \{F(y \leq t|w = d, S = 1)P(w = d|S = 1) + \\ & F[y(d) \leq t|w \neq d, S = 1]P(w \neq d|S = 1)\}P(S = 1) + \\ & F[y(d) \leq t|S = 0]P(S = 0). \end{aligned} \tag{A.8}$$

Suppose that $F[y(d) \leq t|w \neq d, S = 1] = F[y(d) \leq t|S = 0] = 0$. Then

$$F[y(d) \leq t] = F(y \leq t|w = d, S = 1)P(w = d|S = 1)P(S = 1) < \alpha. \tag{A.9}$$

Hence, $Q_\alpha[y(d)] > t$.

4. $r^{NA,NA}(\alpha, d)$ is the largest lower bound.

Let $1-\alpha < P(w=d|S=1)P(S=1) \Longrightarrow P(w\neq d|S=1)P(S=1)+P(S=0)<\alpha$, so that $r^{NA,NA}(\alpha,d)$ is finite. For any $\lambda > 0$,

$$
\begin{aligned}
&F[y(d) \leq r^{NA,NA}(\alpha,d)+\lambda] = F[y(d) \leq r^{NA,NA}(\alpha,d)+\lambda|S=1]P(S=1)+ \\
&F[y(d) \leq r^{NA,NA}(\alpha,d)+\lambda|S=0]P(S=0) = \\
&\{F[y \leq r^{NA,NA}(\alpha,d)+\lambda|w=d,S=1]P(w=d|S=1)+ \\
&F[y(d) \leq r^{NA,NA}(\alpha,d)+\lambda|w\neq d,S=1]P(w\neq d|S=1)\}P(S=1)+ \\
&F[y(d) \leq r^{NA,NA}(\alpha,d)+\lambda|S=0]P(S=0).
\end{aligned}
\tag{A.10}
$$

Suppose that $F[y(d) \leq r^{NA,NA}(\alpha,d)+\lambda|w\neq d,S=1] = F[y(d) \leq r^{NA,NA}(\alpha,d)+\lambda|S=0] = 1$, as is possible in the absence of prior information. Then, the definition of $r^{NA,NA}(\alpha,d)$ implies that

$$
\begin{aligned}
&F[y(d) \leq r^{NA,NA}(\alpha,d)+\lambda] = \\
&\{F[y \leq r^{NA,NA}(\alpha,d)+\lambda|w=d,S=1]P(w=d|S=1)+ \\
&P(w\neq d|S=1)\}P(S=1)+P(S=0) \geq \alpha \Longrightarrow \\
&F[y(d) \leq r^{NA,NA}(\alpha,d)+\lambda] \geq 1-\left\{\frac{(1-\alpha)}{P(w=d|S=1)P(S=1)}\right\}
\end{aligned}
\tag{A.11}
$$

Hence, $Q_\alpha[y(d)] \leq r^{NA,NA}(\alpha,d)+\lambda$.

Suppose now that $1-\alpha \geq P(w=d|S=1)P(S=1) \Longrightarrow P(w\neq d|S=1)P(S=1)+P(S=0) \geq \alpha$, so that $r^{NA,NA}(\alpha,d) = min(\, y^{INFS},\, y^{INFN})$. For any finite t, and supposing that $F[y(d) \leq t|w\neq d,S=1] = F[y(d) \leq t|S=1] = 1$, (A.10) implies that

$$
\begin{aligned}
F[y(d) \leq t] = [&F(y \leq t|w=d,S=1)P(w=d|S=1)+ \\
&P(w\neq d|S=1)]P(S=1)+P(S=0) \geq \alpha.
\end{aligned}
\tag{A.12}
$$

Hence, $Q_\alpha[y(d)] \leq t$.

A.2 Monotone Treatment Response

We now show that the weaker stochastic dominance assumption (14) produces the same MTR bounds as the stronger assumption (13) used by Manski (1997). Looking first at the bounds on $E[y(d)|\, S=1]$ and starting from the NA,NA bounds in (9), we note that both (13) and (14) imply that $E(y|w<d,S=1) \leq E[y(d)|w<d,S=1]$. Hence, $E(y|w<d,S=1)$ can be used in (9) as a lower bound for the counterfactual term $E[y(d)|w<d,S=1]$ instead of y^{INFS}. Similarly, (13) and (14) imply that $E(y|w>d,S=1) \geq E[y(d)|w>d,S=1]$. Hence, $E(y|w>d,S=1)$ can be used in (9) as an upper bound for the counterfactual term $E[y(d)|w>d,S=1]$ instead of y^{SUPS}.

Hence, both (13) and (14) lead to the MTR,NA bounds on $E[y(d)]$ shown in (15). The bounds on $E[y(d)|\, S=1]$ are the same as those derived by Manski (1997) for the case of no sample selection.

Turning now to the bounds on $F[y(d)|S=1]$, and starting from the NA bounds in (10), we note that both (13) and (14) imply that $F(y|w<d,S=1) \geq F[y(d)|w<d,S=1]$. Hence, $F(y|w<d,S=1)$ can be used in (10) as an upper bound for the counterfactual term $F[y(d)|w<d,S=1]$ instead of 1. Similarly, both (13) and (14) imply that $F(y|w>d,S=1) \leq F[y(d)|w>d,S=1]$. Hence, $F(y|w>d,S=1)$ can be used in (10) as a lower bound for the counterfactual term $F[y(d)|w>d,S=1]$ instead of 0. Hence, both (13) and (14) lead to bounds on $F[y(d)]$ under MTR,NA that are equal to

$$\begin{gathered} F(y|w\geq d,S=1)P(w\geq d|S=1)P(S=1) \\ \leq F[y(d)] \leq \\ [F(y|w\leq d,S=1)P(w\leq d|S=1)+P(w>d|S=1)]P(S=1)+P(S=0) \end{gathered} \tag{A.13}$$

We note that the bounds on $F[y(d)|\, S=1]$ are the same as those derived by Giustinelli (2011) for the case of no sample selection.

Proof of Proposition 2. Starting from (A.13) the proof proceeds in the same way as the proof of Proposition 1.

1. $s^{\mathrm{MTR,NA}}(\alpha, d)$ is an upper bound.

   The MTR,NA bounds on $F[y(d)]$ in (A.13) imply that

$$F[y(d)\leq \tilde{y}|w\geq d,S=1]P(w\geq d|S=1)P(S=1)\geq \alpha \implies F[y(d)\leq \tilde{y}]\geq \alpha. \tag{A.144}$$

The premise of (A.14) is empty if $P(w\geq d|S=1)P(S=1)<\alpha$. Suppose that $P(w\geq d|S=1)P(S=1)\geq \alpha$. Then the definition of $s^{MTR,NA}(\alpha,d)$ states that

$$s^{MTR,NA}(\alpha,d) \equiv min\,\tilde{y}\colon F[y(d)\leq \tilde{y}|w\geq d,S=1]\geq \alpha/[P(w\geq d|S=1)P(S=1)]. \tag{A.15}$$

It follows that $F[y(d)\leq s^{MTR,NA}(\alpha,d)]\geq \alpha$. Hence $Q_\alpha[y(d)]\leq s^{MTR,NA}(\alpha,d)$.

This result can be understood from the fact that the upper bound on $Q_\alpha[y(d)]$ is found by reversing the MTR,NA lower bound on $F[y(d)]$ in (A.13). Hence, it is equal to the value $\tilde{y}$ that solves the equation $F(\tilde{y}|w\geq d,S=1)P(w\geq d|S=1)P(S=1)=\alpha \Longrightarrow F(\tilde{y}|w\geq d,S=1)=\frac{\alpha}{P(w\geq d|S=1)P(S=1)}=\beta$. Hence, $\tilde{y}$ is equal to the β^{th} quantile of $F(y|w\geq d,S=1)$, and thus it must be the case that $\beta\leq 1 \Longrightarrow P(w\geq d|S=1)P(S=1)\geq \alpha$.

2. $r^{MTR,NA}(\alpha, d)$ is a lower bound.

The MTR,NA bounds on $F[y(d)]$ in (A.13) imply that

$$\{F[y(d) \leq \tilde{y} | w \leq d, S = 1]P(w \leq \tilde{d} | S = 1) + P(w > d | S = 1)\}P(S = 1) + P(S = 0) < \alpha \Longrightarrow F[y(d) \leq \tilde{y}] < \alpha. \quad \text{(A.16)}$$

Given that $P(w \leq d | S = 1) = 1 - P(w > d | S = 1)$ and $P(S = 0) = 1 - P(S = 1)$, (A.16) may be rewritten as

$$F[y(d) \leq \tilde{y} | w \leq d, S = 1] < 1 - \frac{(1-\alpha)}{P(w \leq d | S=1)P(S=1)} \Longrightarrow F[y(d) \leq \tilde{y}] < \alpha. \quad \text{(A.17)}$$

The premise of (A.17) is empty if $1 - \alpha \geq P(w \leq d | S = 1)P(S = 1) \Longrightarrow P(w > d | S = 1)P(S = 1) + P(S = 0) \geq \alpha$. Suppose that $1 - \alpha < P(w \leq d | S = 1)P(S = 1) \Longrightarrow P(w > d | S = 1)P(S = 1) + P(S = 0) < \alpha$. Then the definition of $r^{MTR,NA}(\alpha, d)$ states that

$$r^{MTR,NA}(\alpha, d) \equiv \min \tilde{y}: F[y(d) \leq \tilde{y} | w \leq d, S = 1] \geq 1 - \left\{\frac{(1 - \alpha)}{P(w \leq d | S = 1)P(S = 1)}\right\}. \quad \text{(A.18)}$$

It follows that, for all $\eta > 0$, $F[y(d) \leq r^{MTR,NA}(\alpha, d) - \eta] < \alpha$. Hence $r^{MTR,NA}(\alpha, d) \leq Q_\alpha[y(d)]$.

This result can be understood from the fact that the lower bound on $Q_\alpha[y(d)]$ is found by reversing the MTR,NA upper bound on $F[y(d)]$ in (A.13). Hence, it is the value $\tilde{y}$ that solves the equation $[F(\tilde{y} | w \leq d, S = 1)P(w \leq d | S = 1) + P(w > d | S = 1)]P(S = 1) + P(S = 0) = \alpha \Longrightarrow F(\tilde{y} | w \leq d, S = 1) = 1 - \frac{1-\alpha}{P(w \leq d | S=1)P(S=1)} = \beta$. Hence, $\tilde{y}$ is equal to the β^{th} quantile of $F(y | w \leq d, S = 1)$, and thus it must be the case that $\beta > 0 \Longrightarrow 1 - \alpha < P(w \leq d | S = 1)P(S = 1) \Longrightarrow P(w > d | S = 1)P(S = 1) + P(S = 0) < \alpha$.

3. $s^{MTR,NA}(\alpha, d)$ is the least upper bound.

Let $P(w \geq d | S = 1)P(S = 1) \geq \alpha$, so that $s^{MTR,NA}(\alpha, d)$ is finite. For any $\lambda > 0$,

$$\begin{aligned} F[y(d) \leq s^{MTR,NA}(\alpha, d) - \lambda] = {} & F[y(d) \leq s^{MTR,NA}(\alpha, d) - \lambda | S = 1]P(S = 1) + \\ & F[y(d) \leq s^{MTR,NA}(\alpha, d) - \lambda | S = 0]P(S = 0) = \\ & \{F[y(d) \leq s^{MTR,NA}(\alpha, d) - \lambda | w < d, S = 1]P(w < d | S = 1) + \\ & F[y(d) \leq s^{MTR,NA}(\alpha, d) - \lambda | w \geq d, S = 1]P(w \geq d | S = 1)\}P(S = 1) + \\ & F[y(d) \leq s^{MTR,NA}(\alpha, d) - \lambda | S = 0]P(S = 0). \end{aligned} \quad \text{(A.19)}$$

Suppose that $F[y(d) \leq s^{MTR,NA}(\alpha,d) - \lambda | w < d, S = 1] = F[y(d) \leq s^{MTR,NA}(\alpha,d) - \lambda | S = 0] = 0$, as is possible in the absence of prior information. Then, the definition of $s^{MTR,NA}(\alpha,d)$ implies that

$$\begin{aligned} &F[y(d) \leq s^{MTR,NA}(\alpha,d) - \lambda] = \\ &\quad F[y \leq s^{MTR,NA}(\alpha,d) - \lambda | w \geq d, S = 1] P(w \geq d | S = 1) P(S = 1) < \alpha. \end{aligned} \tag{A.20}$$

Hence, $Q_\alpha[y(d)] > s^{MTR,NA}(\alpha,d) - \lambda$.

Suppose now that $P(w \geq d | S = 1) P(S = 1) < \alpha$, so that $s^{MTR,NA}(\alpha,d) = max(\, y^{SUPS},\, y^{SUPN})$. For any finite t,

$$\begin{aligned} &F[y(d) \leq t] = \; F[y(d) \leq t | S = 1] P(S = 1) + F[y(d) \leq t | S = 0] P(S = 0) = \\ &\quad \{F(y \leq t | w \geq d, S = 1) P(w \geq d | S = 1) + \\ &\quad F[y(d) \leq t | w < d, S = 1] P(w < d | S = 1)\} P(S = 1) + \\ &\quad F[y(d) \leq t | S = 0] P(S = 0). \end{aligned} \tag{A.21}$$

Suppose that $F[y(d) \leq t | w < d, S = 1] = F[y(d) \leq t | S = 0] = 0$. Then

$$F[y(d) \leq t] = F(y \leq t | w \geq d, S = 1) P(w \geq d | S = 1) P(S = 1) < \alpha. \tag{A.22}$$

Hence, $Q_\alpha[y(d)] > t$.

4. $r^{MTR,NA}(\alpha, d)$ is the largest lower bound.

Let $1 - \alpha < P(w \leq d | S = 1) P(S = 1) \Longrightarrow P(w > d | S = 1) P(S = 1) + P(S = 0) < \alpha$, so that $r^{MTR,NA}(\alpha,d)$ is finite. For any $\lambda > 0$,

$$\begin{aligned} &F[y(d) \leq r^{MTR,NA}(\alpha,d) + \lambda] = \; F[y(d) \leq r^{MTR,NA}(\alpha,d) + \lambda | S = 1] P(S = 1) + \\ &\quad F[y(d) \leq r^{MTR,NA}(\alpha,d) + \lambda | S = 0] P(S = 0) = \\ &\quad \{F[y \leq r^{MTR,NA}(\alpha,d) + \lambda | w \leq d, S = 1] P(w \leq d | S = 1) + \\ &\quad F[y(d) \leq r^{MTR,NA}(\alpha,d) + \lambda | w > d, S = 1] P(w > d | S = 1)\} P(S = 1) + \\ &\quad F[y(d) \leq r^{MTR,NA}(\alpha,d) + \lambda | S = 0] P(S = 0). \end{aligned} \tag{A.23}$$

Suppose that $F[y(d) \leq r^{MTR,NA}(\alpha,d) + \lambda | w > d, S = 1] = F[y(d) \leq r^{MTR,NA}(\alpha,d) + \lambda | S = 0] = 1$, as is possible in the absence of prior information. Then, the definition of $r^{MTR,NA}(\alpha,d)$ implies that

$$\begin{aligned} &F[y(d) \leq r^{MTR,NA}(\alpha,d) + \lambda] = \\ &\quad \{F[y \leq r^{MTR,NA}(\alpha,d) + \lambda | w \leq d, S = 1] P(w \leq d | S = 1) + \end{aligned} \tag{A.24}$$

$P(w > d|S = 1)\}P(S = 1) + P(S = 0) \geq \alpha \implies$

$$F[y(d) \leq r^{MTR,NA}(\alpha, d) + \lambda] \geq 1 - \left\{\frac{(1-\alpha)}{P(w \leq d|S=1)P(S=1)}\right\}.$$

Hence, $Q_\alpha[y(d)] \leq r^{MTR,NA}(\alpha, d) + \lambda$.

Suppose now that $1 - \alpha \geq P(w \leq d|S = 1)P(S = 1) \Longrightarrow P(w > d|S = 1)P(S = 1) + P(S = 0) \geq \alpha$, so that $r^{MTR,NA}(\alpha, d) = min(\, y^{INFS}, \, y^{INFN})$. For any finite t, and supposing that $F[y(d) \leq t|w > d, S = 1] = F[y(d) \leq t|S = 1] = 1$, (A.23) implies that

$$\begin{aligned} F[y(d) \leq t] = [F(y \leq t|w \leq d, S = 1)P(w \leq d|S = 1) + \\ P(w > d|S = 1)]P(S = 1) + P(S = 0) \geq \alpha. \end{aligned} \tag{A.25}$$

Hence, $Q_\alpha[y(d)] \leq t$.

Proof of Lemma 1. Given that $F[y(d)|w = d] = F[y(d)|w = d, S = 0]P(S = 0) + F[y(d)|w = d, S = 1]P(S = 1)$, the stochastic dominance in (24) that holds in each of the two subsamples holds also for the whole sample. That is, $\forall\, d$, and $\forall\, d_1, d_2 \in D$ such that $d_2 > d_1$,

$$F[y(d_2)|w = d] \leq F[y(d_1)|w = d]. \tag{A.26}$$

Since (A.14) holds $\forall\, d$, it also holds unconditionally, that is

$$F[y(d_2)] \leq F[y(d_1)]. \tag{A.27}$$

Both means and quantiles are parameters that respect stochastic dominance, and thus (A.27) implies that both the $ATE(d_2, d_1)$ and the α-$QTE(d_2, d_1)$ are bounded below by zero.

A.3 Monotone Treatment Selection

We first show that the MTS assumption in (18) produces the bounds on $E[y(d)|\, S = 1]$ in (19). Starting from the NA,NA bounds in (9), we note that (18) implies that $E(y|w = d, S = 1) \leq E[y(d)|w > d, S = 1]$. Hence, $E(y|w = d, S = 1)$ can be used in (9) as a lower bound for the counterfactual term $E[y(d)|w > d, S = 1]$ instead of y^{INFS}. Similarly, (18) implies that $E(y|w = d, S = 1) \geq E[y(d)|w < d, S = 1]$. Hence, $E[y|w = d, S = 1]$ can be used in (9) as an upper bound for the counterfactual term $E[y(d)|w < d, S = 1]$ instead of y^{SUPS}. Hence, (18) leads to the MTS bounds on $E[y(d)]$ shown in (19). The MTS bounds on $E[y(d)|\, S = 1]$ are the same as the ones derived by MP for the case of no sample selection.

Turning now to the bounds on $F[y(d)|S=1]$, and starting from the NA bounds in (10), we note that (18) implies that $F(y|w=d,S=1) \leq F[y(d)|w<d,S=1]$. Hence, $F(y|w=d,S=1)$ can be used in (10) as a lower bound for the counterfactual term $F[y(d)|w<d,S=1]$ instead of 0. Similarly, (18) implies that $F(y|w=d,S=1) \geq F[y(d)|w>d,S=1]$. Hence, $F(y|w=d,S=1)$ can be used in (10) as an upper bound for the counterfactual term $F[y(d)|w>d,S=1]$ instead of 1. Hence, (18) lead to bounds on $F[y(d)]$ under MTS that are equal to

$$\begin{gathered} F(y|w=d,S=1)P(w\leq d|S=1)P(S=1) \\ \leq F[y(d)] \leq \\ [F(y|w=d,S=1)P(w\geq d|S=1)+P(w<d|S=1)]P(S=1)+P(S=0) \end{gathered} \tag{A.28}$$

The MTS bounds on $F[y(d)|\,S=1]$ are the same as the ones derived by Giustinelli (2011) for the case of no sample selection.

Proof of Proposition 3. Starting from (A.28) the proof proceeds in the same way as the proof of Proposition 1.

1. $s^{\mathrm{MTS,NA}}(\alpha, d)$ is an upper bound.

   The MTS,NA bounds on $F[y(d)]$ in (A.28) imply that

$$F[y(d)\leq\tilde{y}|w=d,S=1]P(w\leq d|S=1)P(S=1)\geq\alpha \implies F[y(d)\leq\tilde{y}]\geq\alpha. \tag{A.29}$$

The premise of (A.29) is empty if $P(w\leq d|S=1)P(S=1)<\alpha$. Suppose that $P(w\leq d|S=1)P(S=1)\geq\alpha$. Then the definition of $s^{MTS,NA}(\alpha,d)$ states that

$$s^{MTS,NA}(\alpha,d)\equiv min\tilde{y}: F[y(d)\leq\tilde{y}|w=d,S=1]\geq\alpha/[P(w\leq d|S=1)P(S=1)]. \tag{A.30}$$

It follows that $F[y(d)\leq s^{MTS,NA}(\alpha,d)]\geq\alpha$. Hence $Q_\alpha[y(d)]\leq s^{MTS,NA}(\alpha,d)$.

This result can be understood from the fact that the upper bound on $Q_\alpha[y(d)]$ is found by reversing the MTS,NA lower bound on $F[y(d)]$ in (A.28). Hence, it is equal to the value $\tilde{y}$ that solves the equation $F(\tilde{y}|w=d,S=1)P(w\leq d|S=1)P(S=1)=\alpha \Longrightarrow F(\tilde{y}|w=d,S=1)=\frac{\alpha}{P(w\leq d|S=1)P(S=1)}=\beta$. Hence, $\tilde{y}$ is equal to the β^{th} quantile of $F(y|w=d,S=1)$, and thus it must be the case that $\beta\leq 1 \Longrightarrow P(w\leq d|S=1)P(S=1)\geq\alpha$.

2. $r^{MTS,NA}(\alpha, d)$ is a lower bound.

The MTS,NA bounds on $F[y(d)]$ in (A.28) imply that

$$\{F[y(d)\leq\tilde{y}|w=d,S=1]P(w\geq d|S=1)+P(w<d|S=1)\}P(S=1)+ P(S=0)<\alpha\Longrightarrow F[y(d)\leq\tilde{y}]<\alpha. \tag{A.31}$$

Given that $P(w\geq d|S=1)=1-P(w<d|S=1)$ and $P(S=0)=1-P(S=1)$, (A.31) may be rewritten as

$$F[y(d)\leq\tilde{y}|w\geq d,S=1]<1-\frac{(1-\alpha)}{P(w\geq d|S=1)P(S=1)}\Longrightarrow F[y(d)\leq\tilde{y}]<\alpha. \tag{A.32}$$

The premise of (A.32) is empty if $1-\alpha\geq P(w\geq d|S=1)P(S=1)\Longrightarrow P(w<d|S=1)P(S=1)+P(S=0)\geq\alpha$. Suppose that $1-\alpha<P(w\geq d|S=1)P(S=1)\Longrightarrow P(w<d|S=1)P(S=1)+P(S=0)<\alpha$. Then the definition of $r^{MTS,NA}(\alpha,d)$ states that

$$r^{MTS,NA}(\alpha,d)\equiv\min\tilde{y}\colon F[y(d)\leq\tilde{y}|w\geq d,S=1]\geq 1-\left\{\frac{(1-\alpha)}{P(w\geq d|S=1)P(S=1)}\right\}. \tag{A.33}$$

It follows that, for all $\eta>0$, $F[y(d)\leq r^{MTS,NA}(\alpha,d)-\eta]<\alpha$. Hence $r^{MTS,NA}(\alpha,d)\leq Q_\alpha[y(d)]$.

This result can be understood from the fact that the lower bound on $Q_\alpha[y(d)]$ is found by reversing the MTS,NA upper bound on $F[y(d)]$ in (A.28). Hence, it is the value $\tilde{y}$ that solves the equation $[F(\tilde{y}|w=d,S=1)P(w\geq d|S=1)+P(w<d|S=1)]P(S=1)+P(S=0)=\alpha\Longrightarrow F(\tilde{y}|w=d,S=1)=1-\frac{1-\alpha}{P(w\geq d|S=1)P(S=1)}=\beta$. Hence, $\tilde{y}$ is equal to the β^{th} quantile of $F(y|w=d,S=1)$, and thus it must be the case that $\beta>0\Longrightarrow 1-\alpha<P(w\geq d|S=1)P(S=1)\Longrightarrow P(w<d|S=1)P(S=1)+P(S=0)<\alpha$.

3. $s^{MTS,NA}(\alpha, d)$ is the least upper bound.

Let $P(w\leq d|S=1)P(S=1)\geq\alpha$, so that $s^{MTS,NA}(\alpha,d)$ is finite. For any $\lambda>0$,

$$\begin{aligned}
F[y(d)\leq s^{MTS,NA}(\alpha,d)-\lambda]=\ & F[y(d)\leq s^{MTS,NA}(\alpha,d)-\lambda|S=1]P(S=1)+\\
& F[y(d)\leq s^{MTS,NA}(\alpha,d)-\lambda|S=0]P(S=0)=\\
& \{F[y(d)\leq s^{MTS,NA}(\alpha,d)-\lambda|w\leq d,S=1]P(w\leq d|S=1)+\\
& F[y(d)\leq s^{MTS,NA}(\alpha,d)-\lambda|w>d,S=1]P(w>d|S=1)\}P(S=1)+\\
& F[y(d)\leq s^{MTS,NA}(\alpha,d)-\lambda|S=0]P(S=0).
\end{aligned} \tag{A.34}$$

Suppose that $F[y(d)\leq s^{MTS,NA}(\alpha,d)-\lambda|w>d,S=1]=F[y(d)\leq s^{MTS,NA}(\alpha,d)-\lambda|S=0]=0$, as is possible in the absence of prior information. Suppose also that $F[y(d)\leq s^{MTS,NA}(\alpha,d)-$

$\lambda|w < d, S = 1] = F[y(d) \leq s^{MTS,NA}(\alpha, d) - \lambda|w = d, S = 1]$, as is possible under MTS. Then, the definition of $s^{MTS,NA}(\alpha, d)$ implies that

$$\begin{aligned} &F[y(d) \leq s^{MTS,NA}(\alpha, d) - \lambda] = \\ &\quad F[y \leq s^{MTS,NA}(\alpha, d) - \lambda|w \leq d, S = 1]P(w \leq d|S = 1)P(S = 1) = \\ &\quad F[y \leq s^{MTS,NA}(\alpha, d) - \lambda|w = d, S = 1]P(w \leq d|S = 1)P(S = 1) < \alpha. \end{aligned} \tag{A.35}$$

Hence, $Q_\alpha[y(d)] > s^{MTS,NA}(\alpha, d) - \lambda$.

Suppose now that $P(w \leq d|S = 1)P(S = 1) < \alpha$, so that $s^{MTS,NA}(\alpha, d) = max(\, y^{SUPS},\, y^{SUPN})$. For any finite t,

$$\begin{aligned} &F[y(d) \leq t] = F[y(d) \leq t|S = 1]P(S = 1) + F[y(d) \leq t|S = 0]P(S = 0) = \\ &\quad \{F(y \leq t|w \leq d, S = 1)P(w \leq d|S = 1) + \\ &\quad F[y(d) \leq t|w > d, S = 1]P(w > d|S = 1)\}P(S = 1) + \\ &\quad F[y(d) \leq t|S = 0]P(S = 0). \end{aligned} \tag{A.36}$$

Suppose that $F[y(d) \leq t|w > d, S = 1] = F[y(d) \leq t|S = 0] = 0$, as is possible in the absence of prior information. Then

$$F[y(d) \leq t] = F(y \leq t|w \leq d, S = 1)P(w \leq d|S = 1)P(S = 1) < \alpha. \tag{A.37}$$

Hence, $Q_\alpha[y(d)] > t$.

4. $r^{MTS,NA}(\alpha, d)$ is the largest lower bound.

Let $1 - \alpha < P(w \geq d|S = 1)P(S = 1) \Longrightarrow P(w < d|S = 1)P(S = 1) + P(S = 0) < \alpha$, so that $r^{MTS,NA}(\alpha, d)$ is finite. For any $\lambda > 0$,

$$\begin{aligned} &F[y(d) \leq r^{MTS,NA}(\alpha, d) + \lambda] = F[y(d) \leq r^{MTS,NA}(\alpha, d) + \lambda|S = 1]P(S = 1) + \\ &\quad F[y(d) \leq r^{MTS,NA}(\alpha, d) + \lambda|S = 0]P(S = 0) = \\ &\quad \{F[y \leq r^{MTS,NA}(\alpha, d) + \lambda|w \geq d, S = 1]P(w \geq d|S = 1) + \\ &\quad F[y(d) \leq r^{MTS,NA}(\alpha, d) + \lambda|w < d, S = 1]P(w < d|S = 1)\}P(S = 1) + \\ &\quad F[y(d) \leq r^{MTS,NA}(\alpha, d) + \lambda|S = 0]P(S = 0). \end{aligned} \tag{A.38}$$

Suppose that $F[y(d) \leq r^{MTS,NA}(\alpha, d) + \lambda|w < d, S = 1] = F[y(d) \leq r^{MTS,NA}(\alpha, d) + \lambda|S = 0] = 1$, as is possible in the absence of prior information. Suppose also that $F[y(d) \leq r^{MTS,NA}(\alpha, d) - \lambda|w > d, S = 1] = F[y(d) \leq r^{MTS,NA}(\alpha, d) - \lambda|w = d, S = 1]$, as is possible under MTS. Then, the definition of $r^{MTS,NA}(\alpha, d)$ implies that

$$F[y(d) \leq r^{MTS,NA}(\alpha, d) + \lambda] = \tag{A.39}$$

$$\{F[y \le r^{MTS,NA}(\alpha,d)+\lambda | w \ge d, S=1]P(w \ge d|S=1) +$$
$$P(w<d|S=1)\}P(S=1)+P(S=0) =$$
$$\{F[y \le r^{MTS,NA}(\alpha,d)+\lambda | w = d, S=1]P(w \ge d|S=1) +$$
$$P(w<d|S=1)\}P(S=1)+P(S=0) \ge \alpha \implies$$
$$F[y(d) \le r^{MTS,NA}(\alpha,d)+\lambda] \ge 1-\left\{\frac{(1-\alpha)}{P(w \ge d|S=1)P(S=1)}\right\}.$$

Hence, $Q_\alpha[y(d)] \le r^{MTS,NA}(\alpha,d)+\lambda$.

Suppose now that $1-\alpha \ge P(w \ge d|S=1)P(S=1) \Longrightarrow P(w<d|S=1)P(S=1)+P(S=0) \ge \alpha$, so that $r^{MTS,NA}(\alpha,d) = min(\, y^{INFS}, y^{INFN})$. Suppose that for any finite t, $F[y(d) \le t | w<d, S=1] = F[y(d) \le t|S=0] = 1$, as is possible in the absence of prior information. Then, (A.38) implies that

$$\begin{aligned} F[y(d) \le t] = [&F(y \le t | w \ge d, S=1)P(w \ge d|S=1) + \\ &P(w<d|S=1)]P(S=1)+P(S=0) \ge \alpha. \end{aligned} \tag{A.40}$$

Hence, $Q_\alpha[y(d)] \le t$.

When combining MTR and MTS for the selected subsample, and using the results for the $E[y(d)]$ in (15) and (19), we obtain the following bounds on $E[y(d)]$:

$$\begin{aligned} &[E(y|w \le d, S=1)P(w \le d|S=1) + \\ &\quad E(y|w=d,S=1)P(w>d|S=1)]P(S=1)+y^{INFN}P(S=0) \\ &\qquad \le E[y(d)] \le \\ &[E(y|w=d,S=1)P(w<d|S=1) + \\ &\quad E(y|w \ge d, S=1)P(w \ge d|S=1)]P(S=1)+y^{SUPN}P(S=0). \end{aligned} \tag{A.41}$$

We note that the bounds on $E[y(d)|\, S=1]$ under MTR+MTS are the same as those derived in MP for the case of no sample selection.

Turning now to the bounds on $F[y(d)]$, when combining MTR and MTS for the selected subsample, and using the results for the $F[y(d)]$ in (A.13) and (A.16), we obtain the following bounds on $F[y(d)]$:

$$\begin{aligned} &[F(y|w=d,S=1)P(w<d|S=1)+F(y|w \ge d,S=1)P(w \ge d|S=1)]P(S=1) \\ &\qquad \le F[y(d)] \le \\ &[F(y|w \le d,S=1)P(w \le d|S=1) + \\ &\quad F(y|w=d,S=1)P(w>d|S=1)]P(S=1)+P(S=0). \end{aligned} \tag{A.42}$$

We note that the bounds on $F[y(d)|\, S = 1]$ under MTR+MTS are the same as those derived in Giustinelli (2011, p. 795) for the case of no sample selection.

As Giustinelli (2011) points out, the combination of MTR+MTS produces bounds on $F[y(d)|S = 1]$ that are linear combinations of observed cumulative distributions, and thus cannot be inverted analytically to derive bounds on $Q_\alpha[y(d)|S = 1]$. Clearly, the same is true when adding the bounds on $F[y(d)|S = 0]$ to produce the bounds on $F[y(d)]$ in (A.18), and thus bounds on $Q_\alpha[y(d)]$ are calculated by numerical inversion of the bounds on $F[y(d)]$.

A.4 Stochastic Dominance

We now discuss why using the SD assumption narrows the upper bound of the identification range of all quantiles of the distribution of potential outcomes $F[y(d)]$, and why its identification power is much stronger at the upper quantiles. In our discussion, we use MTR+MTS for the selected subsample, and NA for the non-selected one. This implies that, as discussed in Appendix A.3 above, there are no unobserved terms in the lower bound $SLB_F^{MTR+MTS}$ of the distribution $F[y(d)|S = 1]$. In other words, the only unobserved terms of the lower bound $LB_F^{MTR+MTS,NA}$ on $F[y(d)]$ come from the lower bound NLB_F^{NA} on $F[y(d)|S = 0]$. This way the effect of the SD assumption becomes clearer, as its operation on $LB_F^{MTR+MTS,NA}$ (and thus on the upper bounds of the quantiles of $F[y(d)]$) is not affected by any unobserved terms in the lower bound on $F[y(d)|S = 1]$. We note, however, that the reasoning behind the role of the SD assumption goes through for any combination of assumptions on the two subsamples.

Let us take a hypothetical example in which there are 1,000 children in total, 800 of which attend school at grade 6 or above and take the PISA test. The upper bound on the 90th quantile of the population distribution of potential scores (i.e., of $y(d)$) is determined by the lower bound on the population potential score distribution through inversion. Given the potential scores of the 800 children who take the test, the lower bound on the population potential score distribution is given by the score distribution that puts the potential scores of the remaining 200 non-selected children at or above the best potential score among the children that take the test. This is so because putting the 200 potential scores of non-selected children at the top makes the cumulative distribution take small values at low scores. Clearly, this is the best possible distribution of potential outcomes for the children that are not selected, as non-selection is due to either dropping out or lagging behind at school. This, however, implies that the upper bound on the 90th quantile of the population potential score is equal to the 100th best potential score among the children that are not selected. There is no information, however, that one can use to learn something about this score, as no scores are observed

in the non-selected subsample. Hence, this upper bound is not identified, and one has to put it equal to max(y^{SUPS}, y^{SUPN}).

SD rules out this case, as it imposes that non-selected children be stochastically dominated in test scores by the children that take the test, and thus the 90th quantile of the distribution of the potential scores of the latter is also an upper bound on the 90th quantile of the distribution of potential scores of the former, and hence an upper bound on the 90th quantile of the distribution of potential scores for the whole population. Hence, the 90th quantile of the population test score has as an upper bound the 90th quantile of the distribution of potential test scores in the selected subsample, which is equal to the 80th best potential score in that subsample. We thus see that SD has considerably reduced uncertainty at the 90th population quantile, as it reduced its upper bound from max(y^{SUPS}, y^{SUPN}) to a likely quite smaller value.

To give an example, the MTR+MTS,NA upper bound on $Q_{90}[y(d_3)]$, that is, on the 90th quantile of the potential score distribution when the mother's education is tertiary is not identified (the same is true for the corresponding quantile of the potential outcome when mother's education is primary and secondary). Thus, it is set equal to $y^{SUPS} = y^{SUPN}$, that is, to the observed maximum score, which is equal to 741.9 points in PISA. Adding SD reduces it to 557 points.

Using the same reasoning, but now examining what happens at the 10th population quantile, its upper bound without SD would be equal to the 10th quantile of the lower bound on $F[y(d)]$ in (A.42), that is, of $\{F(y|w=d,S=1)P(w<d|S=1)+F(y|w\geq d,S=1)P(w\geq d|S=1)\}P(S=1)$. On the other hand, with SD it would be equal to the 10th quantile of the lower bound on the potential score distribution in the selected subsample $F[y(d)|S=1]$, that is, of $F(y|w=d,S=1)P(w<d|S=1)+F(y|w\geq d,S=1)P(w\geq d|S=1)$. The former distributional lower bound is equal to the latter multiplied by the probability of selection, and thus for a relatively high selection probability (as is the case in our context), the 10th quantiles of these two distributional lower bounds should be relatively close to each other.

To give an example, the MTR+MTS,NA upper bound for children whose mother's education is tertiary ($Q_{10}[y(d_3)]$) is equal to 444.3 points. Adding SD reduces it to 421.2 points.

Finally, as was the case with MTR, imposing the SD assumption in (25) does not imply making any further assumptions on the observability or any other features of the distribution of the realized treatment in the non-selected subsample $G[w|\,S=0]$.

A.5 Bounded Variation

We now discuss how the BVk narrows the identification region by increasing the lower bound on the quantiles of $F[y(d)]$, and especially so for the smaller ones. As with the SD assumption discussed in Appendix A.4, we illustrate the way the BV assumption operates using, without loss of generality,

the MTR+MTS assumption for the selected subsample, and NA for the non-selected one. Hence, there are no unobserved terms in the upper bound $SUB_F^{MTR+MTS}(d)$ on the distribution $F[y(d)|S = 1]$, and the only unobserved terms in the upper bound $UB_F^{MTR+MTS,NA}(d)$ on $F[y(d)]$ come from the upper bound $NUB_F^{NA}(d)$ on $F[y(d)|S = 0]$. This way the effect of the BVk assumption becomes clearer, as its operation on the upper bound on $UB_F^{MTR+MTS}(d)$ (and thus on the lower bounds of the quantiles of $F[y(d)]$) is not affected by any observed terms in the upper bound on $F[y(d)|S = 1]$.

Using similar reasoning as in the case of the SD assumption discussed in Appendix A.4, the lower bound on the 10th quantile is not identified, as it is equal to the 100th best potential score among the non-selected students. Since we know nothing about this subsample, a conservative choice would be to put the lower bound on the 10th population quantile equal to $y^{INFN} = 0$. Using BVk on the other hand, we can reconstruct the upper bound $NUB_F^{BVk}(d)$ in (6) as described in Section 2.4.3, and then solve for the value of y that makes the upper bound on $F[y(d)]$ in (6) equal to .10. This value should be quite higher than $y^{INFN} = 0$.

To give an example, the MTR+MTS,NA+SD lower bound on $Q_{10}[y(d_3)]$ is not identified, and thus is set equal to zero. Adding BV25 raises it to 108.9 points, while adding BV50 raises it to 205.2 points.

On the other hand, the lower bound on the 90th quantile of the population distribution of potential outcomes without imposing BVk (determined by the upper bound on the population distribution of potential outcomes) is the 100th best potential outcome in the selected subsample. Imposing BVk implies substituting the for the term $NUB_F^{BVk}(d) = 1$ in (6) the reconstructed potential outcome distribution of the non-selected subsample $F[y(d)|S = 0]$. We then need to solve for the value of y that makes the upper bound on $F[y(d)]$ in (6) equal to .90. The reconstructed potential outcome distribution, however, is likely close to 1 at this value of y, and thus replacing 1 as a value of $NUB_F^{BVk}(d)$ in (6) through BVk is unlikely to significantly narrow the identification region from above.

To give an example, the MTR+MTS,NA+SD lower bound on $Q_{90}[y(d_3)]$ is equal to 497.6 points. Adding BV25 raises to 498.1 points, while adding BV50 raises to 498.2 points.

Finally, we note again that, as was the case with MTR and SD, imposing the BVk assumption does not imply making any assumptions on the observability or any other features of the distribution of the realized treatment in the non-selected subsample $G[w|\ S = 0]$.

A.6 Monotone Instrumental Variables (MIVs)

We now describe the construction of the MIV bounds on the cumulative distribution $F[y(d)]$, and thus on the $Q_\alpha[y(d)]$. We first note that, as shown in Blundell et al. (2007: 332-333), and given that

the lower (upper) bound on $Q_\alpha[y(d)]$ is determined by the upper (lower) bound on $F[y(d)]$, to get an as large as possible (as small as possible) lower (upper) bound on $Q_\alpha[y(d)]$, one needs to take the minimum (maximum) over the allowed instrument values of the upper (lower) bound on $F[y(d)]$. Hence, for a set of assumptions L on the selected subsample and a set M on the non-selected one, we obtain

$$\max_{z_1 \geq z} LB_F^{L,M}(d|Z = z_1) \leq F[y(d)|Z = z] \leq \min_{z_2 \leq z} UB_F^{L,M}(d|Z = z_2). \tag{A.43}$$

where $LB_F^{L,M}(d|Z = z)$ denotes the lower bound on $F[y(d)|Z = z]$, and $UB_F^{L,M}(d|Z = z)$ denotes the upper bound. As in Blundell et al. (2007), the range of instrument values over which we maximize $LB_F^{L,M}(d|Z = z)$ is one containing instrument values equal to or larger than the one under examination. This is so because, by the MIV assumption in (27), an upper bound on $Q_\alpha[y(d)|z_1]$ is also an upper bound on $Q_\alpha[y(d)|z]$ when $z \leq z_1$, or, alternatively, a lower bound on $F[y(d)|z_1]$ is also a lower bound on $F[y(d)|z]$. In an analogous fashion, to find the upper bound on $F[y(d)|Z = z]$, we minimize $UB_F^{L,M}(d|Z = z)$ over instrument values smaller or equal than the one under examination. This is so because, by the MIV assumption in (27), a lower bound on $Q_\alpha[y(d)|z_2]$ is also a lower bound on $Q_\alpha[y(d)|z]$ when $z \geq z_2$, or, alternatively, an upper bound on $F[y(d)|z_2]$ is also an upper bound on $F[y(d)|z]$.

As is the case with $E[y(d)]$, the bound on $F[y(d)]$ is a weighted average of the bounds in (A.43), with the weights being equal to the probabilities of the instrument values, that is,

$$\begin{aligned} &\sum_z P(Z = z) \max_{z_1 \geq z} LB_F^{L,M}[d|Z = z_1] \\ \leq &\sum_z P(Z = z) F[y(d)|Z = z] = F[y(d)] \leq \\ &\sum_z P(Z = z) \min_{z \geq z_2} UB_F^{L,M}[d|Z = z_2]\,. \end{aligned} \tag{A.44}$$

After calculating the bounds on $F[y(d)]$ in (A.44), one can obtain the bounds on the $Q_\alpha[y(d)]$ by inverting the bounds on $F[y(d)]$.

The operations in (A.43) and (A.44) are illustrated in Fig. A.1 for the case of the upper bound on $F[y(d)]$, which determines the lower bound on $Q_\alpha[y(d)]$. On the horizontal axis we have the values of the outcome Y, while on the vertical axis we have the probability that $F[y(d)]$ and its bounds take values in the support of Y. The α-quantile of the unobserved $F[y(d)]$ is equal to Q_α, as determined by the intersection of the horizontal α-quantile line with $F[y(d)]$.

There are two instrument values, t_1 and t_2, with $t_2 > t_1$. The upper bound conditional on t_1 forms the curve ABICD, while the upper bound conditional on t_2 the curve EBJCF. As t_1 is the smallest possible value, its conditional upper bound remains unchanged after the minimization operation in (A.43). However, to compute the upper bound conditional on t_2 we minimize over the set of all smaller or equal instrument values, that is $\{t_1, t_2\}$. We see that the segment BIC of $UB_F^{L,M}[d|Z = t_1]$ has lower values than the segment BJC of $UB_F^{L,M}[d|Z = t_2]$ and thus the latter bound now forms curve EBICF instead of EBJCF.

Having now determined the upper bounds conditional on t_1 and t_2 through the minimization in (A.43), we need to compute their weighted average using (A.44). The segment BIC is now common to both conditional bounds, and thus the weighted average of the two curves between points B and C coincides with this segment. On the other hand, the weighted average of the two curves to the left of B forms segment GB, while the weighted average to the right C forms segment CH. Hence, the weighted average of the conditional upper bounds in (A.44) is given by the curve GBICH. The lower bound on the α-quantile $Q_\alpha[y(d)]$ is given by Q_α^1.

Importantly, this lower bound is determined simply by inverting the upper bound GBICH on $F[y(d)]$. This upper bound is computed using only (A.43) and (A.44) and without needing to know anything a priori about the lower bound on $Q_\alpha[y(d)]$.

The identifying power of the MIV assumption is clear if one considers what would happen without the minimization operation in (A.43), that is, if we used only the set of assumptions L for the selected subsample and M for the non-selected one. In that case, the weighted average of the two conditional upper bound curves between points B and C would cross the α-quantile horizontal line at a point between J and I, and the resulting lower bound on the quantile would be smaller than Q_α^1.

As in our actual calculations we use more than one instrument, the maximization and minimization operations in (31) and (A.43) take place over vectors of instruments (see de Haan, 2011: 868). Specifically, let us consider two instruments Z^1, and Z^1, and a vector of specific values for them (z_o^1, z_o^2). Maximization operations related to $E[y(d)]$ in (31) are performed over all vectors $(z_j^1, z_k^2,)$ such that $z_j^1 \leq z_o^1$ and, $z_k^2 \leq z_o^2$. Analogously, minimization operations related to $E[y(d)]$ in (31) are performed over all vectors (z_m^1, z_n^2) such that $z_0^1 \leq z_m^1$, and $z_0^2 \leq z_n^2$. These calculations are performed over all possible vectors (z_o^1, z_o^2) of values of the two instruments, as indicated in (32). As for the bounds on $F[y(d)]$, maximization operations (A.43) are performed over all vectors $(z_j^1, z_k^2,)$ such that $z_j^1 \geq z_o^1$ and, $z_k^2 \geq z_o^2$. Analogously, minimization operations in (A.43) are performed over all vectors (z_m^1, z_n^2) such that $z_0^1 \geq z_m^1$, and $z_0^2 \geq z_n^2$. These calculations are performed over all possible vectors (z_o^1, z_o^2) of values of the two instruments, as indicated in (A.44).

Table A.1. Treatment Effects: 25th and 75th Quantiles

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
Assumptions			25th quantile						75th quantile					
Assumptions used for the selected subsample	Assumptions used for the non-selected subsample	MIV	Lower Bound	Upper Bound	Low 95% CI	High 95% CI	Low 90% CI	High 90% CI	Lower Bound	Upper Bound	Low 95% CI	High 95% CI	Low 90% CI	High 90% CI
Panel A. Population treatment effects														
ETS	Random sample selection	No	40.0		36.2	43.7	36.9	43.0	81.2		71.9	86.2	72.9	84.6
NA	NA	No	-392.6	741.9	-400.5	741.9	-398.8	741.9	-741.9	419.0	-741.9	424.6	-741.9	423.4
MTR	NA	No	-47.7	741.9	-49.6	741.9	-49.2	741.9	-57.2	419.0	-60.5	424.6	-59.8	423.4
MTR	MTR	No	0.0	741.9	0.0	741.9	0.0	741.9	0.0	419.0	0.0	424.6	0.0	423.4
MTS	NA	No	-392.6	85.6	-400.5	94.0	-398.8	92.2	-741.9	136.0	-741.9	149.1	-741.9	146.2
MTR+MTS	MTR	No	0.0	85.6	0.0	94.0	0.0	92.2	0.0	136.0	0.0	149.1	0.0	146.2
MTR+MTS	MTR+SD	No	0.0	72.4	0.0	78.6	0.0	77.2	0.0	92.4	0.0	103.4	0.0	101.0
MTR+MTS	MTR+SD+BV25	No	0.0	72.0	0.0	78.1	0.0	76.8	0.0	92.1	0.0	103.1	0.0	100.7
MTR+MTS	MTR+SD+BV50	No	0.0	71.6	0.0	77.6	0.0	76.2	0.0	92.1	0.0	103.1	0.0	100.6
MTR+MTS	MTR+SD	MIV1	0.0	53.1	0.0	58.9	0.0	58.9	0.0	65.2	0.0	69.0	0.0	68.5
MTR+MTS	MTR+SD+BV25	MIV1	0.0	52.9	0.0	58.7	0.0	58.7	0.0	64.9	0.0	68.7	0.0	68.3
MTR+MTS	MTR+SD+BV50	MIV1	0.0	52.7	0.0	58.5	0.0	58.5	0.0	64.8	0.0	68.7	0.0	68.2
MTR+MTS	MTR+SD	MIV2	0.0	33.1	0.0	35.3	0.0	34.6	0.0	43.3	0.0	48.2	0.0	47.2
MTR+MTS	MTR+SD+BV25	MIV2	0.0	32.9	0.0	35.2	0.0	34.3	0.0	42.8	0.0	47.9	0.0	46.9
MTR+MTS	MTR+SD+BV50	MIV2	0.0	32.9	0.0	34.9	0.0	34.3	0.0	42.5	0.0	47.8	0.0	46.8
Panel B. Treatment effects for the selected subsample														
NA	-..-	No	-287.5	655.9	-294.5	655.9	-293.0	655.9	-655.9	397.9	-655.9	402.6	-655.9	401.5
MTR	-..-	No	0.0	655.9	0.0	655.9	0.0	655.9	0.0	397.9	0.0	402.6	0.0	401.5
MTS	-..-	No	-287.5	40.0	-294.5	45.7	-293.0	44.4	-655.9	81.2	-655.9	92.2	-655.9	89.8
MTR+MTS	-..-	No	0.0	40.0	0.0	45.7	0.0	44.4	0.0	81.2	0.0	92.2	0.0	89.8
MTR+MTS	-..-	MIV1	0.0	25.7	0.0	31.2	0.0	31.2	0.0	56.5	0.0	60.5	0.0	60.2
MTR+MTS	-..-	MIV2	3.2	7.8	2.1	10.2	2.3	9.7	0.7	36.1	0.0	40.9	0.0	40.3

Note: See Note to Table 1.

Figure A.1. Using a MIV on the Upper Bound on $F[y(d)]$