Doubly Robust IV Estimation of Local Average Treatment Effects

Selver Derya Uysal*

Department of Economics, CMS

University of Konstanz

July 11, 2011

Abstract

This paper extends the doubly robust methods to the estimation of the Local Average Treatment Effect (LATE) under standard assumptions in the presence of covariates. The proposed method combines regression estimation and weighting estimation of the LATE. The advantage of this combination is that consistency of the LATE is achieved even if the regression estimation or the weighting estimation relies on wrong specifications. The proposed estimation approach is applied to estimate the causal effect of an upper secondary school degree relative to dropping out of high school for those whose high school drop out is induced by a grade retention experience at 10th grade.

JEL classification: $C21, C13, I20$

Keywords: Econometric evaluation, doubly robust estimation, IV, local average treatment effect, propensity score

*Department of Economics, Box D124, University of Konstanz, 78457 Konstanz, Germany. Phone +49-7531-88-2556, Fax -4450, email: selver.uysal@uni-konstanz.de. The author gratefully acknowledges financial support of the German Science Foundation (DFG) through the CMS.

1 Introduction

It has always been a challenge to identify the causal effects of a treatment on an outcome variable from observational studies. Different to experimental studies, in observational studies the selection into the treatment is not random. On the contrary, people usually select themselves into the treatment according to the potential outcome they would obtain with or without the treatment. Due to this selection problem, standard regression methods do not identify the true causal effect.

One possibility to deal with endogenous selection problem is to apply the method of instrumental variables (IV). Although IV has been used in econometric applications since decades, the interpretation of IV estimates in the potential outcome framework is relatively recent. The pioneering papers, which incorporate the IV into the potential outcome framework, are Imbens and Angrist (1994) and Angrist *et al.* (1996). Imbens and Angrist (1994) show under a set of assumptions that the IV estimator identifies the causal effect of the treatment variable on outcome only for the subpopulation whose participation into treatment is induced by the instrument.[1] This causal effect is called Local Average Treatment Effect (LATE). Even though the initial identifying assumptions from Imbens and Angrist (1994) do not include covariates, there is a considerable attempt to extend the LATE concept to include covariates in parametric, semiparametric or nonparametric estimation methods. Abadie (2003), Tan (2006) and Frölich (2007) are the most recent papers, which introduce new estimation methods of the LATE that include covariates.

In this paper, we propose a doubly robust estimation method of the LATE with covariates. The identifying assumptions are similar to those of Abadie (2003), Tan

[1] Under stronger assumptions like homogenous treatment effect, IV identifies the Average Treatment Effect (see for example Abadie (2003)).

(2006) and Frölich (2007). This method is an extension of doubly robust estimation of the Average Treatment Effect (ATE) under the Conditional Independence Assumption (CIA) to estimate the LATE (see Robins *et al.* (1995), Robins and Ritov (1997), Hirano and Imbens (2001), Wooldridge (2007), and Bang and Robins (2005) for further information on doubly robust estimation of the ATE).

The organization of the paper is as follows: Section 2 gives a brief review of the existing methods on estimation of the LATE and proposes the alternative doubly robust method. In Section 3, we present the doubly robustness property of the proposed method in small samples via a simulation study. In Section 4, the proposed estimator is applied to estimate causal effects of upper secondary school graduation on future earnings for the subpopulation whose graduation is induced by a binary instrument grade retention. Finally, Section 5 summarizes the main results and concludes the paper.

2 Estimation of the Local Average Treatment Effect (LATE)

The formal definition of the LATE uses the potential outcomes notation used earlier by Neyman (1923) and Fisher (1935), which became a standard notation in the program evaluation literature after Rubin (1974). D_i shows the binary treatment status for individual i:

$$D_i = \begin{cases} 1, & \text{if the } i^{th} \text{ individual is treated} \\ 0, & \text{otherwise} \end{cases}$$

We define two potential outcomes, Y_{i1} and Y_{i0}, depending on the value of the treatment indicator D_i. For each individual, only one of the potential outcomes is ob-

served. The observed outcome, denoted by Y_i, can be written in terms of the treatment indicator (D_i) and the potential outcomes:

$$Y_i = D_i Y_{i1} + (1 - D_i) Y_{i0} \tag{2.1}$$

The definition of the LATE hinges on the existence of a valid instrument. In this paper, we concentrate on binary instrument, Z_i. Given that the variable Z_i is a valid instrument, we can define the potential treatment status, $\{D_i^z\}$, for the two values of the instrument Z_i. Similar to the observed outcome, we can write the realized treatment status in terms of the instrument Z_i and the potential treatment status:

$$D_i = Z_i D_i^1 + (1 - Z_i) D_i^0. \tag{2.2}$$

According to the relation between the potential treatment status and the binary instrument, we can divide the population into four subpopulations. Following the terminology used by Angrist *et al.* (1996) we demonstrate these four subpopulations in the following table:

Table 1: Partition of the population

	Compliers	Always Takers	Never Takers	Defiers
$Z = 1$	$D^1 = 1$	$D^1 = 1$	$D^1 = 0$	$D^1 = 0$
$Z = 0$	$D^0 = 0$	$D^0 = 1$	$D^0 = 0$	$D^0 = 1$

From the observed data set we cannot identify the group to which an individual belongs since we only observe the pair (D_i, Z_i). For example, if we observe $Z_i = 1$ and $D_i = 1$, we can only say that the individual is either complier or always-taker. Thus, compliers are members of a hypothetically defined subpopulation and cannot be identified from observed data without further assumptions. The LATE is simply

the expected difference between two potential outcomes for the subpopulation of compliers. Formally, we can express the LATE as follows:

$$\tau_{LATE} = E[Y_1 - Y_0 | D^1 > D^0]. \tag{2.3}$$

In order to identify the LATE we use the following assumptions:[2]

A 1 *Conditional Independence of the Instrument:*

$(Y_0, Y_1, D^z) \perp Z|X$ *for each* $z \in \{0, 1\}$.

A 2 *Rank Condition:*

$\Pr[D = 1 | X, Z]$ *is a nontrivial function of* Z*, conditional on* $X = x$.

A 3 *Monotonicity:*

$\Pr[D^1 \geq D^0 | X] = 1$.

A 4 *First Stage:*

$0 < \Pr[Z = 1 | X = x] < 1$ *and* $\Pr[D^1 = 1 | X] > \Pr[D^0 = 1 | X]$.

Assumption A 1 implies that the instrument, Z, is as good as randomly assigned once we condition on the covariates, X. Assumption A 1 also rules out the direct effect of the instrument on the potential outcome. The first two assumption together guarantee that the only effect of the instrument on the outcome is through the treatment variable. Assumption A 4 requires that for any value of X both values of the instrument can be observed. This can be interpreted as a common support assumption. Furthermore, it assures that the instrument and the treatment variable are correlated conditional on the covariates. Assumption A 3 rules out the existence of subpopulations, which are affected by the instrument in an opposite direction. Therefore, the existence of defiers in the population is ruled out.

[2]For the sake of notational simplicity, we drop the running index i.

Imbens and Angrist (1994) show that if Assumptions A 1-3 hold in the absence of covariates, then, the average difference in the outcome variable Y relative to that of the treatment variable D between two instrument groups identifies the LATE:

$$\tau_{LATE} = \frac{E[Y|Z=1] - E[Y|Z=0]}{E[D|Z=1] - E[D|Z=0]}. \tag{2.4}$$

Since it is usually questionable whether we can assume unconditional independence of the instrument, we will concentrate on the identification of LATE conditional on covariates.

2.1 Identification and Existing Methods

In the following part, we summarize the identification results and sketch the proofs in the Appendix.

Theorem 1 *Under Assumptions A 1-A 3 the conditional LATE is identified as*

$$\tau_{LATE}(x) = \mathrm{E}\left[Y_1 - Y_0 | \, X = x, D^1 > D^0\right] = \frac{E[Y|X=x, Z=1] - E[Y|X=x, Z=0]}{E[D|X=x, Z=1] - E[D|X=x, Z=0]} \tag{2.5}$$

Theorem 2 *Under Assumptions A 1-A 4 the unconditional LATE is identified as follows:*

$$\tau_{LATE} = \frac{\mathrm{E}_X\left[\mathrm{E}\left[Y | \, X, Z=1\right] - \mathrm{E}\left[Y | \, X, Z=0\right]\right]}{\mathrm{E}_X\left[\mathrm{E}\left[D | \, X, Z=1\right] - \mathrm{E}\left[D | \, X, Z=0\right]\right]}. \tag{2.6}$$

Theorem 3 *The identification can also be achieved via weighting by the conditional probability of receiving the instrument, which is called the instrument propensity score. Let the probability of receiving the instrument be* $\Pr\left[Z = 1 \, | X\right] = p(X)$*, then,*

the unconditional LATE is identified as follows:

$$\tau_{LATE}^{we} = \mathrm{E}\left[Y_1 - Y_0 | \, X = x, D^1 > D^0\right] = \frac{E\left[\frac{Z}{p(X)}Y\right] - E\left[\frac{1-Z}{1-p(X)}Y\right]}{E\left[\frac{Z}{p(X)}D\right] - E\left[\frac{1-Z}{1-p(X)}D\right]}. \tag{2.7}$$

These identification results show the connection between the LATE and the ATE. The ratio in Equation (2.5) can be seen as a ratio of two conditional ATEs: ATE of Z on Y, $\tau^{Y|Z}(x)$, divided by ATE of Z on D, $\tau^{D|Z}(x)$. The same relation holds also for the unconditional effects. We can rewrite Equations (2.5) and (2.6) as:

$$\begin{aligned}
\tau_{LATE}(x) &= \frac{E[Y|X=x, Z=1] - E[Y|X=x, Z=0]}{E[D|X=x, Z=1] - E[D|X=x, Z=0]} = \frac{\tau^{Y|Z}(x)}{\tau^{D|Z}(x)} \\
\tau_{LATE} &= \frac{\mathrm{E}_X\left[\mathrm{E}\left[Y|\, X, Z=1\right] - \mathrm{E}\left[Y|\, X, Z=0\right]\right]}{\mathrm{E}_X\left[\mathrm{E}\left[D|\, X, Z=1\right] - \mathrm{E}\left[D|\, X, Z=0\right]\right]} = \frac{\mathrm{E}_X\left[\tau^{Y|Z}(x)\right]}{\mathrm{E}_X\left[\tau^{D|Z}(x)\right]} = \frac{\tau_{Y|Z}}{\tau_{D|Z}}.
\end{aligned}$$

Although $\tau^{Y|Z}$ and $\tau^{D|Z}$ are not real treatment effects, the identification problem, however, is similar to the identification problem when estimating the ATE. We can rewrite Eq. 2.1 and the conditional means of the outcome variable by substituting the definition of the potential treatment status as follows:

$$\begin{aligned}
Y_i &= (Z_i D_i^1 + (1-Z_i)D_i^0)Y_{i1} + (1 - (Z_i D_i^1 + (1-Z_i)D_i^0))Y_{i0} \\
\mathrm{E}\left[Y_i|\, Z_i = 1\right] &= \mathrm{E}\left[Y^1\right] = D_i^1 Y_{i1} + (1 - D_i^1)Y_{i0} \\
\mathrm{E}\left[Y_i|\, Z_i = 0\right] &= \mathrm{E}\left[Y^0\right] = D_i^0 Y_{i1} + (1 - D_i^0)Y_{i0}.
\end{aligned}$$

As it is clear from the above equations, these means are not identified from observed data because one of the potential outcomes on the right-hand side is always missing. The estimation problem boils down to the estimation of four unconditional means with a missing data problem: two unconditional means of potential outcomes with respect to the instrument Z ($\mathrm{E}\left[Y^1\right]$ and $\mathrm{E}\left[Y^0\right]$) and two unconditional means of the potential treatment variable with respect to the instrument Z ($\mathrm{E}\left[D^1\right]$ and $\mathrm{E}\left[D^0\right]$).

The goal is to estimate these unconditional means consistently to get consistent estimates of the LATE. Using the connection between the LATE and the ATE, we can borrow estimation methods of the ATE under CIA to get estimators of the LATE (see Hirano *et al.* (2003), Imbens (2004), Wooldridge (2002) Ch. 18 for further information on estimation of the ATE). For example, Frölich (2007) provides a nonparametric estimator for the estimation of the LATE with covariates, which is basically the ratio of two matching estimators of ATEs of Z on Y and D. Tan (2006), on the other hand, derives a parametric regression method for the LATE estimation, where he identifies $\mathrm{E}\left[Y|\, X, Z\right]$ by $\mathrm{E}\left[Y|\, D, X, Z\right]$ and $\mathrm{E}\left[D|\, X, Z\right]$. Another obvious possibility is to use parametric regression adjustment to get two ATEs. The conditional mean functions $E[Y|X, Z = z]$ and $E[D|X, Z = z]$ for $z \in \{0, 1\}$ can be estimated using the individuals who receive the instrument ($z = 1$) and do not receive the instrument ($z = 0$) separately if the assumptions listed previously are fulfilled. Assume $m_z(X, \beta_z)$ and $\mu_z(X, \alpha_z)$ are models for $E[Y^z|X]$ and $E[D^z|X]$ for $z = 0, 1$ respectively. The coefficients can be estimated by regression methods. The M-Estimator based representation of the regression estimation can be written as:

$$
\begin{aligned}
\{\hat{\beta}_1\} &= arg\min \frac{1}{N}\sum_i Z_i q_1^y(Y_i, X_i; \beta_1) \\
\{\hat{\alpha}_1\} &= arg\min \frac{1}{N}\sum_i Z_i q_1^d(D_i, X_i; \alpha_1) \\
\{\hat{\beta}_0\} &= arg\min \frac{1}{N}\sum_i (1 - Z_i) q_0^y(Y_i, X_i; \beta_0) \\
\{\hat{\alpha}_0\} &= arg\min \frac{1}{N}\sum_i (1 - Z_i) q_0^d(D_i, X_i; \alpha_0).
\end{aligned}
$$

where $q^y(\cdot)$ and $q^d(\cdot)$ are the objective functions. If $m_z(\cdot)$ and $\mu_z(\cdot)$ are correct specifications of the conditional means, with consistent, $\sqrt{N}$-asymptotically normal estimator of β_z and α_z for $z = 0, 1$, we get consistent estimators of $\mathrm{E}\left[Y^z\right]$ and $\mathrm{E}\left[D^z\right]$

and as a result consistent estimators of $\tau^{Y|Z}$ and $\tau^{D|Z}$. Using the resulting estimated parameters, the LATE can be estimated by:

$$\hat{\tau}^{reg}_{LATE} = \frac{\frac{1}{N}\sum_i^N (m_1(X_i, \hat{\beta}_1) - m_0(X_i, \hat{\beta}_0))}{\frac{1}{N}\sum_i^N (\mu_1(X_i, \hat{\alpha}_1) - \mu_0(X_i, \hat{\alpha}_0))} = \frac{\hat{\tau}^{Y|Z}}{\hat{\tau}^{D|Z}}. \tag{2.8}$$

The consistency of the estimators of $\mathrm{E}[Y^z]$ and $\mathrm{E}[D^z]$, however, hinges upon the correct specification of the models for the conditional mean functions, $m_z(X, \beta_z)$ and $\mu_z(X, \alpha_z)$. As an example, we assume a linear specification for the conditional mean function, i.e., $m_1(X_i, \beta_1) = X_i'\beta_1$. A consistent estimation of the unconditional mean requires that $\mathrm{E}[Z_i(Y_i^1 - X_i'\beta_1)]$ is equal to zero. By law of iterated expectations, we can write the following equality:

$$\begin{aligned} \mathrm{E}\left[Z_i(Y_i^1 - X_i'\beta_1)\right] &= \mathrm{E}\left[\mathrm{E}\left[Z_i(Y_i^1 - X_i'\beta_1)\middle| X_i\right]\right] \\ &= \mathrm{E}\left[\mathrm{E}\left[Z_i\middle| X_i\right]\left(\mathrm{E}\left[Y_i^1\middle| X_i\right] - X_i'\beta\right)\right]. \end{aligned}$$

This expectation is equal to zero only if $\mathrm{E}[Y_i^1| X_i] = X_i'\beta_1$, i.e., if the conditional mean is correctly specified. The relation is the same for the other three terms necessary for the estimation of the LATE.

Following the identification result in Theorem 3, the LATE can be estimated by replacing the $p(X)$ in Eq. (2.7) by its parametric estimate and the expectations by sample means as proposed by Tan (2006) or as a ratio of two propensity score matching estimators as proposed by Frölich (2007). The parametric weighting estimator

of the LATE is given by:

$$\begin{aligned}\hat{\tau}^{we}_{LATE} &= \frac{N^{-1}\sum_i\left[\frac{Z_i}{\hat{p}(X_i;\hat{\gamma})}Y_i\right]-N^{-1}\sum_i\left[\frac{1-Z_i}{1-\hat{p}(X_i;\hat{\gamma})}Y_i\right]}{N^{-1}\sum_i\left[\frac{Z_i}{\hat{p}(X_i;\hat{\gamma})}D_i\right]-N^{-1}\sum_i\left[\frac{1-Z_i}{1-\hat{p}(X_i;\hat{\gamma})}D_i\right]} \\ &= \frac{N^{-1}\sum_i\frac{[Z_i-\hat{p}(X_i;\hat{\gamma})]Y_i}{\hat{p}(X_i;\hat{\gamma})[1-\hat{p}(X_i;\hat{\gamma})]}}{N^{-1}\sum_i\frac{[Z_i-\hat{p}(X_i;\hat{\gamma})]D_i}{\hat{p}(X_i;\hat{\gamma})[1-\hat{p}(X_i;\hat{\gamma})]}},\end{aligned} \tag{2.9}$$

where $\hat{p}(X_i;\hat{\gamma})$ is the estimated probability of receiving the instrument. This method is an extension of propensity score weighting estimation of the ATE. The weighting by the inverse of the probability, however, estimates the unconditional mean consistently as long as the probability of receiving the instrument is correctly specified. This can be seen from the proof of Theorem 3 in Appendix A. To get from the second equality to the third, it is necessary that $\mathrm{E}[Z|X] = p(X)$. Therefore, if $p(X)$ is wrongly specified, the weighting does not recover the unconditional mean.

The asymptotic distribution of the above mentioned estimators of the LATE can be derived easily for a known joint asymptotic distribution of the ATE estimators $\hat{\tau}^{Y|Z}$ and $\hat{\tau}^{D|Z}$, which satisfy:

$$\sqrt{N}\left(\begin{pmatrix}\hat{\tau}^{Y|Z}\\ \hat{\tau}^{D|Z}\end{pmatrix}-\begin{pmatrix}\tau^{Y|Z}\\ \tau^{D|Z}\end{pmatrix}\right)\xrightarrow{d}N(0,\Omega),$$

where Ω is variance-covariance matrix. Since the LATE is simply the ratio of these two estimators, we can derive the asymptotic distribution of the LATE estimator by the Delta Method:

$$\sqrt{N}(\hat{\tau}_{LATE}-\tau_{LATE})\xrightarrow{d}N\left(0,\left(\frac{1}{\tau^{D|Z}}\right)^2V_{Y|Z}+\left(\frac{\tau^{Y|Z}}{(\tau^{D|Z})^2}\right)^2V_{D|Z}-2\frac{\tau^{Y|Z}}{(\tau^{D|Z})^3}\operatorname{Cov}\left(\tau^{Y|Z},\tau^{D|Z}\right)\right). \tag{2.10}$$

In order to estimate the variance of the LATE estimator, we replace the unknown

population parameters $\tau^{Y|Z}$, $\tau^{D|Z}$ and their variances with their estimates.

2.2 Doubly Robust Estimation

In this paper, we propose a doubly robust parametric estimation method similar to the doubly robust estimator of the ATE as in Wooldridge (2007) (see also Robins *et al.* (2008)). It shares the same property as the doubly robust estimator of the LATE proposed by Tan (2006).[3] It is essentially the weighted regression of the four conditional means in Equation 2.8 with the weights $\frac{1}{\hat{P}(Z=1|X=x)}$ for $Z = 1$ and $\frac{1}{1-\hat{P}(Z=1|X=x)}$ for $Z = 0$ in order to estimate β_z and α_z.

$$\begin{aligned}
\{\hat{\beta}_1^w\} &= arg\min \frac{1}{N}\sum_i \frac{Z_i}{p(X_i;\hat{\gamma})} q_1^y(Y_i, X_i; \beta_1) \\
\{\hat{\alpha}_1^w\} &= arg\min \frac{1}{N}\sum_i \frac{Z_i}{p(X_i;\hat{\gamma})} q_1^d(D_i, X_i; \alpha_1) \\
\{\hat{\beta}_0^w\} &= arg\min \frac{1}{N}\sum_i (\frac{1-Z_i}{1-p(X_i;\hat{\gamma})}) q_0^y(Y_i, X_i; \beta_0) \\
\{\hat{\alpha}_0^w\} &= arg\min \frac{1}{N}\sum_i (\frac{1-Z_i}{1-p(X_i;\hat{\gamma})}) q_0^d(D_i, X_i; \alpha_0),
\end{aligned}$$

where $q(\cdot)$ is the objective function and $\hat{\gamma}$ is the estimated parameter vector for the instrument propensity score.

$$\hat{\tau}_{LATE}^{dr} = \frac{\frac{1}{N}\sum_i m_1(X_i, \hat{\beta}_1^w) - \frac{1}{N}\sum_i m_0(X_i, \hat{\beta}_0^w)}{\frac{1}{N}\sum_i \mu_1(X_i, \hat{\alpha}_1^w) - \frac{1}{N}\sum_i \mu_0(X_i, \hat{\alpha}_0^w)} \tag{2.11}$$

[3] Tan (2006) proposes a different combination of the regression method and weighting method. In order to get the doubly robust estimator of the LATE, the first term in the denominator in Eq. 2.7 is replaced by

$$E\left[\frac{Z}{p(1|x)}Y\right] - E\left[\left(\frac{Z}{p(1|x)} - 1\right) E[Y|X = x, Z = 1]\right]$$

the second term in the denominator is replaced by

$$E\left[\frac{1-Z}{1-p(1|x)}Y\right] - E\left[\left(\frac{1-Z}{1-p(1|x)} - 1\right) E[Y|X = x, Z = 0]\right]$$

and the terms in nominator are replaced similarly.

This doubly robust estimator of the LATE is consistent if the instrument propensity score is correctly specified or the mean functions of the outcome variable and the treatment variable are correctly specified, whereas the consistency of the LATE estimator based on a unweighted regression or inverse instrument propensity score weighting hinges upon the correct specification of the relevant models. In short, it is enough to have one of the methods correct in order to get consistent estimators of the LATE. Given that it is almost impossible to be sure whether a method is correct, this doubly robust method provides some safety in applied work. Depending on which method we are using we need to specify either the instrument propensity score model (Eq. 2.7) or the conditional mean functions of the outcome variable and the treatment variable (Eq. 2.8). Let the $m_z(\cdot)$ and $\mu_z(\cdot)$ be the correct specifications of the conditional means. If we use a weighted regression method with the weights $\frac{1}{P(Z=1|X=x)}$ for $Z=1$ and $\frac{1}{1-P(Z=1|X=x)}$ for $Z=0$ to estimate β_z and α_z, the estimator of the LATE is consistent even if $P(Z=1|X=x)$ is wrongly specified. The LATE estimator in Equation 2.7, however, would not be consistent with wrongly specified instrument propensity score as illustrated above. We can use the previous example with a linear mean function $m_1(X_i,\beta_1) = X_i'\beta_1$ to show the doubly robustness of our estimator. A consistent estimation of the unconditional mean requires that $\mathrm{E}\left[\frac{Z_i}{p(X_i,\gamma)}(Y_i^1 - X_i'\beta_1)\right]$ is equal to zero. By the law of iterated expectations, we can write the following equality:

$$\begin{aligned}
\mathrm{E}\left[\frac{Z_i}{p(X_i,\gamma)}(Y_i^1 - X_i'\beta_1)\right] &= \mathrm{E}\left[\mathrm{E}\left[\frac{Z_i}{p(X_i,\gamma)}(Y_i^1 - X_i'\beta_1)\middle| X_i\right]\right] \\
&= \mathrm{E}\left[\frac{\mathrm{E}\left[Z_i|\, X_i\right]}{p(X_i,\gamma)}\mathrm{E}\left[(Y_i^1 - X_i'\beta_1)\middle|\, X_i\right]\right] \\
&= \mathrm{E}\left[\frac{\mathrm{E}\left[Z_i|\, X_i\right]}{p(X_i,\gamma)}(\mathrm{E}\left[Y_i^1\middle|\, X_i\right] - X_i'\beta_1)\right].
\end{aligned} \tag{2.12}$$

This shows that, even if $p(X_i,\gamma)$ is a wrong specification of $\Pr\left[Z_i = 1\,|X_i\right]$, $E[Y_i^1]$ is consistently estimated as long as $\mathrm{E}\left[Y_i^1|\, X_i\right] = X_i'\beta_1$ holds, i.e., the expecta-

tion is equal to zero. Moreover, with a correctly specified instrument propensity score we can get a consistent estimator of the LATE for certain combinations of $m_z(\cdot)$ and $\mu_z(\cdot)$ even if $m_z(\cdot)$ and $\mu_z(\cdot)$ are wrongly specified. For models satisfying $E[Y^z] = E[m_z(X, \hat{\beta}_z)]$ and $E[D^z] = E[\mu_z(X, \hat{\alpha}_z)]$ although $E[Y^z|X] \neq m_z(X, \hat{\beta}_z)$ and $E[D^z|X] \neq \mu_z(X, \hat{\alpha}_z)$, weighted regression will estimate the LATE consistently. We know that linear, logistic and Poisson regression models satisfy this relation. In our example for a linear model, if $p(X_i, \gamma)$ is the correct specification for $\mathrm{E}\left[Z_i \middle| X_i\right]$, but $E[Y_i^1|X_i] \neq X_i'\beta_1$, by properties of linear model Eq. (2.12) simplifies to:

$$\mathrm{E}\left[\mathrm{E}\left[Y_i^1 \middle| X_i\right] - X_i'\beta_1\right] = \mathrm{E}\left[Y^1\right] - \mathrm{E}\left[X_i'\beta_1\right] = 0$$

even if $\mathrm{E}\left[Y_i^1 \middle| X_i\right] \neq X_i'\beta_1$. Therefore, if we choose the mean functions among these models given that the distributional assumptions are in line with the characteristics of the outcome variables, $\hat{\tau}_{LATE}$ will be a consistent estimator of the LATE, if $P(Z = 1|X = x)$ is correctly specified or $E[Y^z|X]$ and $E[D^z|X]$ are correctly specified.

The doubly robust estimation of the LATE requires the estimation of the conditional mean functions and the instrument propensity score to generate the weights. This can be done by a two step M-Estimation procedure, where the weights are estimated in the first step and used in the second step (see Wooldridge (2002) p. 353- 356 for asymptotic distribution of two step M-Estimators). Another approach can be joint estimation of all parameters in M-Estimation framework. Furthermore, by adding the estimation of $\tau^{Y|Z}$ and $\tau^{D|Z}$ into the joint estimation procedure, we can easily get $\mathrm{Cov}\left(\tau^{Y|Z}, \tau^{Y|Z}\right)$, which we need in order to estimate the variance of the LATE estimator (see Eq. 2.10). Let $\theta = (\beta_1, \beta_0, \alpha_1, \alpha_0, \gamma, \tau^{Y|Z}, \tau^{D|Z})$ and $W = (Y, X, D, Z)$. The estimators can be defined as a solution for the sample

moment equation

$$\frac{1}{N}\sum_{i=1}^{N}\psi(W_i,\hat{\theta}) = 0. \tag{2.13}$$

By standard results for M-Estimation it follows that:

$$\sqrt{N}(\hat{\theta}-\theta) \overset{a}{\sim} N\left(0, A^{-1}VA^{-1}\right) \tag{2.14}$$

where

$$A \equiv E\left[\frac{\partial\psi(W_i,\theta)}{\partial\theta'}\right] \tag{2.15}$$

$$V \equiv V[\psi(W_i,\theta)] = E[\psi(W_i,\theta)\psi(W_i,\theta)']$$

For the estimation of the instrument propensity score, we can use a probit or logit estimation method. For both, the relevant moment function will be the score of the loglikelihood. Depending on the nature of the outcome variable the proper mean function is a generalized linear model with the identity, logit or poisson link function. The model for the mean can be written as follows:

$$m_z(X_i,\beta_z) = g(X_i'\beta_z), \tag{2.16}$$

where $g(\cdot)^{-1}$ is the canonical link function. For a continuous outcome variable the suitable link function is the identity link, whereas for a dichotomous outcome the logit link ($g(a)^{-1} = \ln\left(\frac{a}{1-a}\right)$), $g(a) = \frac{\exp(a)}{1+\exp(a)}$) and for a nonnegative discrete outcome variable the log link ($g(a)^{-1} = \ln(a)$, $g(a) = \exp(a)$) will be suitable. Thus, the natural choice for the mean of a binary treatment indicator is a generalized linear model with the logit link:

$$\mu(X_i;\alpha_z) = \Lambda(X_i'\alpha_z) = \frac{\exp(X_i'\alpha_z)}{1+\exp(X_i'\alpha_z)}.$$

If the instrument propensity score is specified as a logit function $P(Z_i = 1|X_i) =$

$\Lambda(X_i'\gamma) = \frac{\exp(X_i'\gamma)}{1+\exp(X_i'\gamma)}$, the moment functions related to each mean function can be written as follows:

$$
\begin{aligned}
\psi_1(W,\theta) &= \frac{Z}{\Lambda(X\gamma)} X'(Y - m_1(X,\beta_1)) \\
\psi_2(W,\theta) &= \frac{1-Z}{1-\Lambda(X\gamma)} X'(Y - m_0(X,\beta_0)) \\
\psi_3(W,\theta) &= \frac{Z}{\Lambda(X\gamma)} X'(D - \Lambda(X\alpha_1)) \\
\psi_4(W,\theta) &= \frac{1-Z}{1-\Lambda(X\gamma)} X'(D - \Lambda(X\alpha_0)) \\
\psi_5(W,\theta) &= X(Z - \Lambda(X\gamma)) \\
\psi_6(W,\theta) &= m_1(X,\beta_1) - m_0(X,\beta_0) - \tau^{Y|Z} \\
\psi_7(W,\theta) &= \mu_1(X,\alpha_1) - \mu_0(X,\alpha_0) - \tau^{D|Z}.
\end{aligned}
$$

The moment function in Eq. 2.13 can be written in terms of these seven moment functions:

$$
\psi(W,\theta) = \begin{pmatrix} \psi_1(W,\theta) \\ \psi_2(W,\theta) \\ \psi_3(W,\theta) \\ \psi_4(W,\theta) \\ \psi_5(W,\theta) \\ \psi_6(W,\theta) \\ \psi_7(W,\theta) \end{pmatrix}
$$

The M-estimators of $\tau^{Y|Z}$ and $\tau^{Y|Z}$ from the weighted moment functions can be used to estimate the LATE robustly. In the appendix, we provide the variance estimator, which we need to estimate the variance of $\hat{\tau}^{dr}_{LATE}$ based on the asymptotic distribution given in Eq. (2.10). Note that the regression LATE estimator can be calculated in a similar way. The difference is that the fifth moment condition does

not exist and the other moment conditions are only multiplied by Z_i or $(1-Z_i)$, but not weighted by the instrument propensity score.

3 Simulation Study

Assume that $X_{1i}, X_{2i}, Z_i, D_{zi}, Y_{di}$ are generated as follows. First, X_1 and X_2 are normally distributed random variables with unit variances and 0.2 and -0.2 means. Second, logistically distributed error terms ν , ε are drawn. The binary instrument and the potential treatment status are determined by the following rules:

$$Z = \mathbb{1}\{\gamma_0 + \gamma_1 X_1 + \gamma_2 X_2 + \nu > 0\}$$
$$D^1 = \mathbb{1}\{\alpha_{10} + \alpha_{11} X_1 + \alpha_{12} X_2 + \varepsilon > 0\}$$
$$D^0 = \mathbb{1}\{\alpha_{00} + \alpha_{01} X_1 + \alpha_{02} X_2 + \varepsilon > 0\}.$$

where $\gamma = (0.5, -0.5, 0.5)'$, $\alpha_1 = (0.7, -0.2, -0.7)'$ and $\alpha_0 = (-0.3, -0.2, -0.7)'$. To guarantee the monotonicity assumption and the existence of compliers, the coefficients of the X's and the error terms in the potential treatment status are chosen to be equal and the constant α_{10} is chosen larger to be than α_{00}. The potential outcomes are generated according to the following data generating processes (DGPs):

$$Y_1 = \beta_{10} + \beta_{11} X_1 + \beta_{12} X_2 + \epsilon_1$$
$$Y_0 = \beta_{00} + \beta_{01} X_1 + \beta_{02} X_2 + \epsilon_0$$

, where $\beta_1 = (3, 4, 2)'$, $\beta_0 = (1, 5, -1)'$ and (ϵ_1, ϵ_0) are independent vectors of standard normal variables. Finally, we set $D = D^z$ if $Z = z$ and $Y = Y_d$ if $D = d$ for $d, z \in \{0, 1\}$. For the second setting, X_1 and X_2 are drawn from a multivariate normal distribution with the same mean vector as in the first setting, unit variances and 0.4 covariance. The distribution of the other variables are the same.

We compare three different LATE estimators. The first one is estimated as in Equation 2.11, where the coefficients β_z and α_z are estimated by unweighted regression (denoted by REG in the tables). The second method is the parametric weighting method (IPS). The conditional expectation of Z in Equation 2.7 is replaced by its logit estimates. The third method is the doubly robust method proposed in this paper (DR). The coefficients of the mean functions are estimated by weighted regression. To apply these methods, we need three specifications: the instrument propensity score, the treatment propensity score, and the conditional mean of the outcome. We estimate all these three specifications under various scenarios. First, we use correct specifications for all of them. Under correct specifications, all three methods estimate the LATE consistently. Second, we estimate the instrument propensity score using a wrong specification, where X_2 is omitted. In this case, all methods except the weighting method other methods estimate the LATE consistently. This kind of misspecification does not affect the consistency of regression method, since it does not rely on the instrument propensity score. Although the doubly robust method uses the estimated instrument propensity score, the resulting LATE estimator is still consistent, since the conditional mean functions of Y and D are correctly specified. Third, we use wrong specifications for the conditional mean functions of Y and D. We exclude X_2 for these specifications, but we use a correct specification for the instrument propensity score. Under this scenario, the weighting method and the doubly robust method are consistent, whereas the regression method results in an inconsistent LATE estimator. We also consider the cases where only one of the conditional mean functions is wrongly specified, i.e, either the conditional mean function of Y or D is wrongly specified. The results are based on 5000 Monte Carlo samples with sample sizes n =500, 1000 and 4000. We report in the tables mean bias (BIAS) and mean bias relative to the true LATE (MRBIAS) over Monte Carlo

replications. Additionally, standard errors are estimated for each replication according to the asymptotic variance formula given in Section 2 and the mean standard errors (SE) are reported. The standard errors are not informative for the cases where the estimators are inconsistent. Lastly, the mean squared errors (MSE) are given in the tables. Table B 1 and B 2 in Appendix summarize the results from the first setting where the X's are independent and from the second setting where the X's are correlated, respectively.

We clearly observe the doubly robustness property of our proposed method in Table B 1, whereas single methods are severely affected by wrong specifications. If all specifications are correct, shown in Panel A in Table B 1, the regression method is the most efficient method in terms of estimated standard errors and MSE. The doubly robust Method, however, is slightly less efficient due to additional estimation error coming from the estimated instrument propensity score. IPS is the least efficient method. For the smallest sample size, the MSE of the IPS method is twice as large as the MSE of the regression and doubly robust methods. Panel B of the same table shows the results with misspecified instrument propensity score. Clearly, the IPS estimator is biased if the instrument propensity score is misspecified as the theory suggests. The regression method is unaffected by misspecification since it does not require the estimation of instrument propensity score. The doubly robust method, on the other hand, uses the misspecified instrument propensity score, but this does not harm its consistency, since the regression part is correctly specified. If we compare the two consistent methods under this setting, we observe that the efficiency difference is very small. The results in Panel C are for the setting where a correct specification of the instrument propensity score is used but various misspecification combinations in the regression part are present. As expected, the regression method delivers inconsistent estimators under each misspecification scenario. The

consequences of misspecification in terms of bias and efficiency is more severe if both required specifications are wrong (Panel C1). Comparison between Panel C2 and C3 shows that the regression method is less affected by misspecification of the mean function of the treatment indicator than by misspecification of the mean function of the outcome variable. Even though the effect is smaller, the regression method is still inconsistent if only the mean function of the treatment indicator is misspecified.

Table B 2 summarizes the results for correlated X's. The results on the consistency of different methods are the same as before as the theory suggest. The efficiency difference between three methods decreases. The doubly robust method is almost as efficient as the regression method when compared for the cases where the regression method is consistent. When compared to IPS for the cases the IPS method is consistent, the doubly robust method is still more efficient, even though the efficiency difference is smaller.

The settings of both tables are replicated after checking for the common support assumption according to minima and maxima comparison as in the ATE estimation literature under CIA (see Caliendo and Kopeinig (2008) and their references for the discussion on common support). Adapting this criterion, we delete all observations whose instrument propensity score is smaller than the minimum and larger than the maximum in the opposite instrument group. The results suggest that at most 3% of the data is excluded. The small percentage of the off-support observations occurs as a result of our data generating process. Therefore, adjustments to generate a common support do not change the previous Monte Carlo results.[4]

[4]The results are not reported here.

4 Empirical Application

In this section, the causal effect of having an upper secondary school graduation on earnings is investigated for the individuals whose graduation from upper secondary school is instrumented by grade retention. The empirical part uses the data from a longitudinal panel study of 3240 10th grade students attending 121 classes at 68 advanced secondary schools (Gymnasien) in North Rhine-Westphalia (Central Archive for Empirical Social Research (2007), Meulemann (2007), Rauber (2007)). Although the dataset is restricted to students attending upper secondary school in North Rhine-Westphalia, the empirical study is still representative for Germany. One fifth of the German population resides in North Rhine-Westphalia and it is the biggest federal state in terms of population among the 16 federal states in Germany. Furthermore, one forth of the students in Germany is attending school in North Rhine-Westphalia. Besides, the upper secondary schools in Germany serve almost for one half of the total students after primary education (Grundschule).[5] The tests and interviews were conducted at three different points in time: during the 10th grade (1970), at the age of 30 (1984) and at the age of 43 (1997). The 10th grade students were asked questions about their characteristics, school background and relations with parents. They also participated in a psychometric test. The first wave also contains parent and teacher questionnaires. Around 1980, the students' grades were collected from the schools. The last two waves in 1984 and 1997 contain information about the employment and academic history between the last interview and the current one. The sample size was reduced to about 1600 participating individuals at the age 43, which is about 50 percent of the initial sample size.

Estimation of causal effects of schooling on earnings has been an important challenge

[5]The exact numbers can be found on the website of Federal Statistical Office: http://www.destatis.de/.

in empirical economics due to the selection problems (see Card (1999), Card (2001) for related issues). However, here we are interested in average returns of schooling for those who comply with assignment to the treatment mechanism implied by the instrument. Here, the treatment variable is the completion of upper secondary school versus dropping out of upper secondary school after 10th grade, whereas the binary policy instrument is grade retention at 10th grade. We examine the effect of the upper secondary school diploma on the earnings for those whose high school degree is affected by the policy instrument grade retention. In education research it has been shown that the grade retention is one of the most important determinants of high school drop out (see Eide and Showalter (2001), Alexander *et al.* (2003), Jacob and Lefgren (2002) among others). The choice of instrumental variable estimation is not because we cannot identify the ATE, but because we are especially interested in the causal effects of schooling on those individuals who can be seen as at risk of dropping out of high school due to implementation of grade retention.

The earnings variable is constructed as net monthly income divided by average hours worked per month for all individuals that worked at least once between 1984 and 1997.[6] To claim that the identifying assumptions in Section 2 hold approximately, it is necessary to include confounding variables which affect the potential treatment, potential outcomes as well as the instrumental variable. As in the ATE estimation under CIA assumption, a rich set of confounding variables is very crucial. Moreover, we have to be certain that the confounding variables are not affected by the treatment or the instrumental variable. It is important to note that all the variables are measured before instrument and treatment status are observed. Therefore, it is less likely that the covariates are affected by treatment or instrumental status.

[6]Net monthly income in the data set is inflation adjusted. Average hours worked are measured by actual and not by contractually specified hours. Average hours worked per month are calculated as weekly hours times four.

Given that many recent papers like Heckman *et al.* (2006), Carneiro *et al.* (2007), **?**, Wichert and Pohlmeier (2010), Uysal and Pohlmeier (2010) and Rauber (2007) provide empirical evidence on the importance of noncognitive and cognitive skills in determining different outcomes such as school performance, earnings, labor force participation, and job finding success, it is advantageous that our dataset gives us the possibility to measure certain dimensions of noncognitive skills and cognitive skills besides the usual control variables. To account for noncognitive skills, we use the information attribution of success as in Rauber (2007) and Flossmann (2010). Whether a person attributes her success to internal factors, such as diligence and ability, or to external factors, such as family or luck, is closely related to Rotter's (1966) concept of the locus of control. Individuals who attribute success to internal factors have higher self-esteem, and therefore higher noncognitive skills, whereas individuals who attribute success to external factors do not take responsibility for their lives and blame other for their failures, thus, they are more likely to have lower noncognitive skills. A measure of intelligence, IQ, is also included to control for the cognitive skills of the students. The variable which accounts for cognitive skills is the sum of correctly solved questions of a standard psychometric Intelligence Structure Test (IST), which was administered in the class-room in the 10th grade. Besides noncognitive and cognitive skills, we use the information on the desire of further studies to control for motivation. All variables used in the study are listed in Table B 3.

After selecting the explanatory variables, all observations with missing information for any of the explanatory variables, outcome variables, treatment variable and instrument are dropped. This decreases the sample size to 1552. Table B 4 summarizes the descriptive statistics of the variables for the entire sample. It also reports means and standard deviations of the variables in the sample by the treatment variable

upper secondary school diploma, D, and the binary policy instrument grade retention, Z. 67% of the sample has a high school diploma and 71% of the sample did not repeat 10th grade. The proportion of grade retainees who hold a high school diploma is 30%, whereas 71% of the non-retainees earned a high school diploma. Relative to the high school graduates, those without a high school degree earn less. On average, the high school graduates have higher IQ scores and are younger. The average age of the mothers is larger for the high school graduates than for non-graduates. The parents of high school graduates are on average more educated, earn more and show more interest on their children's school outcomes than those of non-graduates The high school graduates attribute their success less to their diligence, luck and family than the non-graduates. The other measures of noncognitive skills, ABIL and ASTUTE, are not statistically different for high school graduates and non-graduates. The variable WISH which measures the motivation of the students differ also significantly between the two groups. The last two columns of Table B 4 show means and standard deviations of the covariates for those who have not been retained and for those who have been. Comparison of the averages of the covariates for nonretained and retained gives similar numbers to those found between high school graduates and non-graduates except for household income and noncognitive measures. The household income does not differ significantly for the retainees and non-retainees. The individuals who repeated the 10th grade attribute their success to astuteness and family more than those who did not repeat. The other measures of noncognitive ability do not differ between the two groups.

With the help of Table B 4, we can compute some simple estimators, which are only consistent if the treatment or the instrument can be assumed to be independent of potential outcomes. If graduation from upper secondary school were independent of potential earnings, we could estimate the ATE of graduation from upper secondary

school as the difference of the sample means by treatment status. This comparison estimates the ATE as 0.22 (0.02) for log-wages. This naive estimator, however, is likely to be biased since the individuals select themselves into treatment according to their potential outcomes. If grade retention were a valid instrument in the absence of the covariates, we could estimate the LATE by Equation 2.4, which gives 0.30 (0.08). However, it is hard to believe that the potential outcomes are independent of the instrument without controlling for covariates. The significant differences in the averages of covariates also support the assumption that unconditional independence is difficult to claim. Therefore, we proceed with our doubly robust estimation method which relies on identification assumptions conditional on the covariates. For model selection purposes, we also estimate the LATE by the unweighted regression method and simple inverse instrument propensity score weighting.

The estimation of the instrument propensity score is carried out by using a logit regression on the covariates. Table B 5 reports the logit estimates. From the results, we can conclude that females are less likely to be retained. IQ has a decreasing effect on the probability of being retained on average. Having a young mother increases the probability of being retained. The other variable, which is highly significant is the willingness to pursue higher education (WISH). The probability of being retained is lower for students who are planning to pursue higher education. The variable ABIL is slightly significant and positive, where as the variable FAMILY is significantly negative. This results has the fairly intuitive interpretation that the internal attribution, therefore higher noncognitive skills, increases the probability of being not retained, while external attribution, lower noncognitive skills, increases the probability of being retained. In general, these results coincide with previous findings in education literature.

As in the estimation of the ATE, we can evaluate the common support assumption (Assumption A 4) by comparing the distributions (histograms) of the estimated instrument propensity scores by instrumental variable as suggested in Lechner (2010). Figure C 1 shows that there is no common support problem.[7]

In order to apply the doubly robust or the regression method, we also specify the conditional mean function of the outcome variable, LNWAGE, and the conditional mean function of the treatment variable, D. Since our dependent variable is continuous, the identity link is chosen. For the binary treatment variable, we use the logit link function. Table 2 presents the LATE estimates and the standard errors based on the different methods.[8]

Table 2: LATE Estimates

Method	Estimate	Std. Err
Doubly Robust	0.38	0.085
Weighting	0.40	0.125
Regression	0.37	0.089

All three LATE estimates are statistically significant, and they are larger than the mean comparisons by treatment status and the simple Wald estimate. The doubly

[7]The observations shown as off support in Figure C 1 are those individuals who belong to the group $Z = 1$ and whose probability of receiving the instrument is larger than the maximum probability of receiving the instrument for the group (Z=0). Since there are only 9 observations in off support, we did not drop them, but it is illustrated in the graph for the sake of completeness.

[8]The similarity of the estimates is not a coincidence. We used the doubly robust property of our proposed method as an informal model selection criterion. If all the specifications are correct, all three methods provide consistent estimates of the LATE. The differences, in estimates, however, can be an indication of wrong specification. We perform the estimation with various specifications and pick the specification which gives the closest estimates by different methods. For some other specifications the estimates differ substantially among methods.

robust LATE estimate is 38%. This means that individuals whose upper secondary school graduation is induced by grade retention earn on average 38% more if they graduate from upper secondary school. This estimate is larger than the standard OLS estimates of returns to schooling and even larger than other IV estimates of returns to schooling (see Flossmann and Pohlmeier (2006), Ichino and Winter-Ebmer (2004), Ichino and Winter-Ebmer (1999)). However, this does not mean that our method overestimates the LATE. By definition, the LATE varies with the instruments chosen, simply because the causal effect is identified for different subpopulations depending on the instrument. In this study, we examine average causal effects of obtaining an upper secondary school degree on earnings relative to the state of dropping out of upper secondary school for those whose dropping out of the school is induced by repeating a grade. Our result shows that grade retention as an educational policy instrument does worsen the future income of retainees indirectly.

5 Conclusion

In this paper, we propose a combination of weighting and regression methods of the LATE estimation which is doubly robust. The weighting method alone requires a correct specification of the instrument propensity score to get consistent LATE estimators, whereas the regression method requires a correct specification of the outcome and treatment mean functions. To apply the proposed method, we need to specify both sets of models, but in order to achieve consistency it is sufficient to have one set correctly specified.

In a small Monte Carlo study, we demonstrate the doubly robustness property in small samples. The theoretical results are valid also for small samples. The weight-

ing method is the least efficient one. In the estimation of the ATE, the weighting method has shown to be less efficient since the variance is highly affected by too large and too small propensity score estimates. According to our simulation results, the same problem appears in the estimation of the LATE. The doubly robust method, on the other hand, is slightly less efficient than the regression. Despite the fact that using the estimated weights causes additional estimation error, the efficiency loss is ignorable.

In our empirical application, the causal effect of having an upper secondary school graduation on earnings is investigated for those individuals whose graduation from upper secondary school is instrumented by grade retention. We use the doubly robust property of our proposed estimator as a model selection criteria and choose the specification, which delivers similar results for the different methods. The LATE estimates with the proposed instrument is larger than the standard OLS estimates of returns to schooling and even larger than other IV estimates of returns to schooling. The implication of this result is that the individuals whose graduation from upper secondary school is induced by grade retention are affected by the treatment more than other subpopulations.

References

Abadie, A. (2003): "Semiparametric Instrumental Variable Estimation of Treatment Response Models," *Journal of Econometrics*, 113, 231–263.

Alexander, K. L., D. R. Entwisle, and S. L. Dauber (2003): *On the Success of Failure: A Reassessment of the Effects of Retention in the Primary Grades*. Cambridge University Press, Cambridge.

Amthauer, R. (1953): *Intelligenz-Struktur-Test*. Verlag fr Psychologie, Dr. C.J. Hogrefe, Gttingen, 2. erweiterte auflage edn.

Angrist, J., G. Imbens, and D. Rubin (1996): "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association*, 91, 444 – 472.

Bang, H., and J. M. Robins (2005): "Doubly Robust Estimation in Missing Data and Causal Inference Models," *Biometrics*, 61, 962–972.

CAESR (2007): *Dataset Gymnasiastenstudie*. Central Archive for Empirical Social Research, Köln.

Caliendo, M., and S. Kopeinig (2008): "Some Practical Guidance for the Implementation of Propensity Score Matching," *Journal of Economic Surveys*, 22, 31–72.

Card, D. (1999): "The Causal Effect of Education and Earnings," *Handbook of Labour Economics*, 3, 1801–1863.

——— (2001): "Estimating the Returns to Schooling: Progress on Some Persistent Econometric Problems," *Econometrica*, 69, 1127–1160.

Carneiro, P., C. Crawford, and A. Goodman (2007): "Which Skills Matter?," in *Practice Makes Perfect: The Importance of Practical Learning*, ed. by D. Kehoe, pp. 22–38. Social Markets Foundation, London.

Eide, E. R., and M. H. Showalter (2001): "The Effect of Grade Retention on Education and Labor Market Outcomes," *Economics of Education Review*, 20, 563–576.

Fisher, R. A. (1935): *Design of Experiments*. Oliver and Boyd, London.

Flossmann, A. (2010): "Accounting for missing data in M-estimation: a general matching approach," *Empirical Economics*, 38, 85–117.

Flossmann, A., and W. Pohlmeier (2006): "Causal Returns to Education: A Survey in Empirical Evidence for Germany," *Journal of Economics and Statistics*, 226, 6–23.

Frölich, M. (2007): "Nonparametric IV Estimation of Local Average Treatment Effect with Covariates," *Journal of Econometrics*, 139, 3575.

Heckman, J. J., J. Stixrud, and S. Urzua (2006): "The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior," *Journal of Labor Economics*, 24, 411 – 482.

Hirano, K., and G. Imbens (2001): "Estimation of Causal Effects using Propensity Score Weighting: An Application to Data on Right Heart Catheterization," *Health Services and Outcomes Research Methodology*, 2, 259–278.

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

Ichino, A., and R. Winter-Ebmer (1999): "Lower and upper bounds of returns to schooling: An exercise in IV estimation with different instruments," *European Economic Review*, 43(4-6), 889 – 901.

——— (2004): "The Long-Run Educational Cost of World War II," *Journal of Labor Economics*, 22(1), 57–87.

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

Imbens, G., and J. Angrist (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62, 467 – 476.

Jacob, B. A., and L. Lefgren (2002): "Remedial Education and Student Achievement: A Regression- Discontinuity Analysis," NBER Working Paper 8918, Cambridge, MA: National Bureau of Economic Research.

Lechner, M. (2010): "A note on the common support problem in applied evaluation studies," *Annales d'conomie et de Statistique*, forthcoming.

Meulemann, H. (2007): "Projektbericht zur Vorlage bei der DFG," Az.Me 577/7-1; -2.

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

Rauber, M. (2007): "Noncognitive Skills and Success in Life: The Importance of Motivation and Self- Regulation," Discussion Paper 07.

Robins, J., M. Sued, Q. Lei-gomez, A. Rotnitzky, F. Ciencias, and E. Naturales (2008): "Comment: Performance of Double-Robust Estimators When Inverse Probability Weights Are Highly Variable," *Statistical Science*, 22.

Robins, J. M., and Y. Ritov (1997): "Toward a Curse of Dimensionality Appropriate (CODA) Asymptotic Theory for Semi-Parametric Models," *Statistics in Medicine*, 16, 285–319.

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

Rotter, J. (1966): "Generalized Expectancies for Internal versus External Control of Reinforcement," *Psychological Monographs*, 80, 1–28.

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

Tan, Z. (2006): "Regression and Weighting Methods for Causal Inference Using Instrumental Variables," *Journal of the American Statistical Association*, 101, 1607–1618.

Uysal, S. D., and W. Pohlmeier (2010): "Unemployment Duration and Personality Traits," Working Paper, University of Konstanz.

Wichert, L., and W. Pohlmeier (2010): "Female Labor Force Participation and the Big Five," ZEW Discussion Paper No. 10-003.

Wooldridge, J. M. (2002): *Econometric Analysis of Cross Section and Panel Data*. MIT Press, Cambridge, MA.

——— (2007): "Inverse Probability Weighted Estimation for General Missing Data Problems," *Journal of Econometrics*, 141, 1281 – 1301.

A Appendix

A.1 Proofs

[9]

Proof 1

$$
\begin{aligned}
& \mathrm{E}\left[Y \middle| X, Z=1\right] - \mathrm{E}\left[Y \middle| X, Z=0\right] = \\
= \; & \mathrm{E}\left[DY_1 + (1-D)Y_0 \middle| X, Z=1\right] - \mathrm{E}\left[DY_1 + (1-D)Y_0 \middle| X, Z=0\right] \\
= \; & \mathrm{E}\left[(ZD^1 + (1-Z)D^0)Y_1 + (1 - ZD^1 - (1-Z)D^0)Y_0 \middle| X, Z=1\right] \\
- \; & \mathrm{E}\left[(ZD^1 + (1-Z)D^0)Y_1 + (1 - ZD^1 - (1-Z)D^0)Y_0 \middle| X, Z=0\right] \\
= \; & \mathrm{E}\left[D^1Y_1 - D^1Y_0 \middle| X, Z=1\right] - \mathrm{E}\left[D^0Y_1 - D^0Y_0 \middle| X, Z=0\right] \\
= \; & \mathrm{E}\left[D^1(Y_1 - Y_0) \middle| X\right] - \mathrm{E}\left[D^0(Y_1 - Y_0) \middle| X\right] \\
= \; & \mathrm{E}\left[(Y_1 - Y_0)(D^1 - D^0) \middle| X\right]
\end{aligned}
$$

The first three equations follow from the definition of the potential outcome and the potential treatment status. The fourth equation follows from Assumption A 1.

$$
\begin{aligned}
\mathrm{E}\left[(Y_1 - Y_0)(D^1 - D^0) \middle| X\right] &= \mathrm{E}\left[Y_1 - Y_0 \middle| X, D^1 - D^0 = 1\right] \Pr\left[D^1 - D^0 = 1 \,|X\right] \\
&- \mathrm{E}\left[Y_1 - Y_0 \middle| X, D^1 - D^0 = 1\right] \Pr\left[D^1 - D^0 = -1 \,|X\right] \\
&= \mathrm{E}\left[Y_1 - Y_0 \middle| X, D^1 - D^0 = 1\right] \Pr\left[D^1 - D^0 = 1 \,|X\right]
\end{aligned}
$$

Therefore,

$$
\tau_{LATE}(x) = \mathrm{E}\left[Y_1 - Y_0 \middle| X, D^1 > D^0\right] = \frac{\mathrm{E}\left[Y \middle| X, Z=1\right] - \mathrm{E}\left[Y \middle| X, Z=0\right]}{\Pr\left[D^1 > D^0 \,|X\right]}.
$$

Due to Assumption A 3, the second term in the first equation is equal to zero. Moreover, since $\mathrm{E}\left[D \middle| X, Z=0\right] = \Pr\left[D=1 \,|X, Z=0\right] = P[always\ takers|X] + P[defiers|X]$ *and* $\mathrm{E}\left[D \middle| X, Z=1\right] = \Pr\left[D=1 \,|X, Z=1\right] = P[always\ takers|X] + P[compliers|X]$*, the relative size of the subpopulation of compliers is identified as:*

$$
\Pr\left[D^1 > D^0 \,|X\right] = \mathrm{E}\left[D \middle| X, Z=1\right] - \mathrm{E}\left[D \middle| X, Z=0\right].
$$

Therefore,

$$
\tau_{LATE}(x) = \mathrm{E}\left[Y_1 - Y_0 \middle| X, D^1 - D^0 = 1\right] = \frac{\mathrm{E}\left[Y \middle| X, Z=1\right] - \mathrm{E}\left[Y \middle| X, Z=0\right]}{\mathrm{E}\left[D \middle| X, Z=1\right] - \mathrm{E}\left[D \middle| X, Z=0\right]}.
$$

Proof 2 *The conditional LATE has to be averaged over X in the compliers subpopulation*

[9] We follow for the proofs mainly Frölich (2007).

in order to get the unconditional LATE

$$\begin{aligned}
\tau_{LATE} &= E_{X|D^1>D^0}[\tau_{LATE}(x)] \\
&= \int \tau_{LATE}(x) f(x|D^1 > D^0)dx \\
&= \int \tau_{LATE}(x) \frac{\Pr\left[D^1 > D^0 \,|X = x\right]}{\Pr\left[D^1 > D^0\right]} f(x)dx
\end{aligned}$$

where the last equation follows from Bayes' Rule. We insert the definition of the conditional LATE in the above equation:

$$\begin{aligned}
\tau_{LATE} &= \int \frac{\mathrm{E}\left[Y|\, X = x, Z = 1\right] - \mathrm{E}\left[Y|\, X = x, Z = 0\right]}{\mathrm{E}\left[D|\, X = x, Z = 1\right] - \mathrm{E}\left[D|\, X = x, Z = 0\right]} \frac{\Pr\left[D^1 > D^0 \,|X = x\right]}{\Pr\left[D^1 > D^0\right]} f(x)dx \\
&= \frac{\int \mathrm{E}\left[Y|\, X = x, Z = 1\right] - \mathrm{E}\left[Y|\, X = x, Z = 0\right] f(x)dx}{\Pr\left[D^1 > D^0\right]} \\
&= \frac{\mathrm{E}_X\left[\mathrm{E}\left[Y|\, X = x, Z = 1\right] - \mathrm{E}\left[Y|\, X = x, Z = 0\right]\right]}{\Pr\left[D^1 > D^0\right]}.
\end{aligned}$$

From the first to the second equation $\Pr\left[D^1 > D^0 \,|X = x\right]$ *and* $(\mathrm{E}\left[D|\, X, Z = 1\right] - \mathrm{E}\left[D|\, X, Z = 0\right])$ *cancel, and* $\Pr\left[D^1 > D^0\right]$ *is taken out of the integral since it is independent of* X*. Note that:*

$$\begin{aligned}
\Pr\left[D^1 > D^0\right] &= \underset{X}{\mathrm{E}}\left[\mathrm{E}\left[D^1 > D^0 \middle|\, X\right]\right] \\
&= \underset{X}{\mathrm{E}}\left[\mathrm{E}\left[D|\, X, Z = 1\right] - \mathrm{E}\left[D|\, X, Z = 0\right]\right].
\end{aligned}$$

Thus, the unconditional LATE is identified as

$$\tau_{LATE} = \frac{\mathrm{E}_X\left[\mathrm{E}\left[Y|\, X, Z = 1\right] - \mathrm{E}\left[Y|\, X, Z = 0\right]\right]}{\mathrm{E}_X\left[\mathrm{E}\left[D|\, X, Z = 1\right] - \mathrm{E}\left[D|\, X, Z = 0\right]\right]}$$

Proof 3

$$\begin{aligned}
\mathrm{E}\left[\frac{Z}{p(X)}Y\right] &= \mathrm{E}\left[\mathrm{E}\left[\frac{Z}{p(X)}Y \middle|\, X\right]\right] \\
&= \mathrm{E}\left[\frac{\mathrm{E}\left[Z|\, X\right]}{p(X)}\mathrm{E}\left[Y|\, X\right]\right] \\
&= \mathrm{E}\left[\mathrm{E}\left[Y|\, X, Z = 1\right]\right]
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{E}\left[\frac{Z}{p(X)}D\right] &= \mathrm{E}\left[\mathrm{E}\left[\frac{Z}{p(X)}D \middle|\, X\right]\right] \\
&= \mathrm{E}\left[\frac{\mathrm{E}\left[Z|\, X\right]}{p(X)}\mathrm{E}\left[D|\, X\right]\right] \\
&= \mathrm{E}\left[\mathrm{E}\left[D|\, X, Z = 1\right]\right]
\end{aligned}$$

Thus,

$$\tau_{LATE}^{we} = \mathrm{E}\left[Y_1 - Y_0 | \, X = x, D^1 > D^0\right] = \frac{E\left[\frac{Z}{p(X)}Y\right] - E\left[\frac{1-Z}{1-p(X)}Y\right]}{E\left[\frac{Z}{p(X)}D\right] - E\left[\frac{1-Z}{1-p(X)}D\right]}$$

A.2 Variance Estimation

In order to estimate the asymptotic variance we replace the unknown parameter vector θ with its estimate $\hat{\theta}^w$ and the expectations with sample means in Eq. (2.14):

$$\begin{aligned}
\hat{V} &= \frac{1}{N}\sum_i \psi(W_i, \hat{\theta}^w)\psi(W_i, \hat{\theta}^w)' \\
\hat{A} &= \frac{1}{N}\sum_i \frac{\partial \psi(W_i, \hat{\theta}^w)}{\partial \theta'} \\
&= \frac{1}{N}\sum_i \begin{pmatrix}
\frac{\partial\psi_1(W_i,\hat{\theta}^w)}{\partial\beta_1'} & \frac{\partial\psi_1(W_i,\hat{\theta}^w)}{\partial\beta_0'} & \frac{\partial\psi_1(W_i,\hat{\theta}^w)}{\partial\alpha_1'} & \frac{\partial\psi_1(W_i,\hat{\theta}^w)}{\partial\alpha_0'} & \frac{\partial\psi_1(W_i,\hat{\theta}^w)}{\partial\gamma'} & \frac{\partial\psi_1(W_i,\hat{\theta}^w)}{\partial\tau^{Y|Z}} & \frac{\partial\psi_1(W_i,\hat{\theta}^w)}{\partial\tau^{D|Z}} \\
\frac{\partial\psi_2(W_i,\hat{\theta}^w)}{\partial\beta_1'} & \frac{\partial\psi_2(W_i,\hat{\theta}^w)}{\partial\beta_0'} & \frac{\partial\psi_2(W_i,\hat{\theta}^w)}{\partial\alpha_1'} & \frac{\partial\psi_2(W_i,\hat{\theta}^w)}{\partial\alpha_0'} & \frac{\partial\psi_2(W_i,\hat{\theta}^w)}{\partial\gamma'} & \frac{\partial\psi_2(W_i,\hat{\theta}^w)}{\partial\tau^{Y|Z}} & \frac{\partial\psi_2(W_i,\hat{\theta}^w)}{\partial\tau^{D|Z}} \\
\frac{\partial\psi_3(W_i,\hat{\theta}^w)}{\partial\beta_1'} & \frac{\partial\psi_3(W_i,\hat{\theta}^w)}{\partial\beta_0'} & \frac{\partial\psi_3(W_i,\hat{\theta}^w)}{\partial\alpha_1'} & \frac{\partial\psi_3(W_i,\hat{\theta}^w)}{\partial\alpha_0'} & \frac{\partial\psi_3(W_i,\hat{\theta}^w)}{\partial\gamma'} & \frac{\partial\psi_3(W_i,\hat{\theta}^w)}{\partial\tau^{Y|Z}} & \frac{\partial\psi_3(W_i,\hat{\theta}^w)}{\partial\tau^{D|Z}} \\
\frac{\partial\psi_4(W_i,\hat{\theta}^w)}{\partial\beta_1'} & \frac{\partial\psi_4(W_i,\hat{\theta}^w)}{\partial\beta_0'} & \frac{\partial\psi_4(W_i,\hat{\theta}^w)}{\partial\alpha_1'} & \frac{\partial\psi_4(W_i,\hat{\theta}^w)}{\partial\alpha_0'} & \frac{\partial\psi_4(W_i,\hat{\theta}^w)}{\partial\gamma'} & \frac{\partial\psi_4(W_i,\hat{\theta}^w)}{\partial\tau^{Y|Z}} & \frac{\partial\psi_4(W_i,\hat{\theta}^w)}{\partial\tau^{D|Z}} \\
\frac{\partial\psi_5(W_i,\hat{\theta}^w)}{\partial\beta_1'} & \frac{\partial\psi_5(W_i,\hat{\theta}^w)}{\partial\beta_0'} & \frac{\partial\psi_5(W_i,\hat{\theta}^w)}{\partial\alpha_1'} & \frac{\partial\psi_5(W_i,\hat{\theta}^w)}{\partial\alpha_0'} & \frac{\partial\psi_5(W_i,\hat{\theta}^w)}{\partial\gamma'} & \frac{\partial\psi_5(W_i,\hat{\theta}^w)}{\partial\tau^{Y|Z}} & \frac{\partial\psi_5(W_i,\hat{\theta}^w)}{\partial\tau^{D|Z}} \\
\frac{\partial\psi_6(W_i,\hat{\theta}^w)}{\partial\beta_1'} & \frac{\partial\psi_6(W_i,\hat{\theta}^w)}{\partial\beta_0'} & \frac{\partial\psi_6(W_i,\hat{\theta}^w)}{\partial\alpha_1'} & \frac{\partial\psi_6(W_i,\hat{\theta}^w)}{\partial\alpha_0'} & \frac{\partial\psi_6(W_i,\hat{\theta}^w)}{\partial\gamma'} & \frac{\partial\psi_6(W_i,\hat{\theta}^w)}{\partial\tau^{Y|Z}} & \frac{\partial\psi_6(W_i,\hat{\theta}^w)}{\partial\tau^{D|Z}} \\
\frac{\partial\psi_7(W_i,\hat{\theta}^w)}{\partial\beta_1'} & \frac{\partial\psi_7(W_i,\hat{\theta}^w)}{\partial\beta_0'} & \frac{\partial\psi_7(W_i,\hat{\theta}^w)}{\partial\alpha_1'} & \frac{\partial\psi_7(W_i,\hat{\theta}^w)}{\partial\alpha_0'} & \frac{\partial\psi_7(W_i,\hat{\theta}^w)}{\partial\gamma'} & \frac{\partial\psi_7(W_i,\hat{\theta}^w)}{\partial\tau^{Y|Z}} & \frac{\partial\psi_7(W_i,\hat{\theta}^w)}{\partial\tau^{D|Z}}
\end{pmatrix} \\
&= \frac{1}{N}\sum_i \begin{pmatrix}
\frac{\partial\psi_1(W_i,\hat{\theta}^w)}{\partial\beta_1'} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \frac{\partial\psi_1(W_i,\hat{\theta}^w)}{\partial\gamma'} & \mathbf{0} & \mathbf{0} \\
\mathbf{0} & \frac{\partial\psi_2(W_i,\hat{\theta}^w)}{\partial\beta_0'} & \mathbf{0} & \mathbf{0} & \frac{\partial\psi_2(W_i,\hat{\theta}^w)}{\partial\gamma'} & \mathbf{0} & \mathbf{0} \\
\mathbf{0} & \mathbf{0} & \frac{\partial\psi_3(W_i,\hat{\theta}^w)}{\partial\alpha_1'} & \mathbf{0} & \frac{\partial\psi_3(W_i,\hat{\theta}^w)}{\partial\gamma'} & \mathbf{0} & \mathbf{0} \\
\mathbf{0} & \mathbf{0} & \mathbf{0} & \frac{\partial\psi_4(W_i,\hat{\theta}^w)}{\partial\alpha_0'} & \frac{\partial\psi_4(W_i,\hat{\theta}^w)}{\partial\gamma'} & \mathbf{0} & \mathbf{0} \\
\mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \frac{\partial\psi_5(W_i,\hat{\theta}^w)}{\partial\gamma'} & \mathbf{0} & \mathbf{0} \\
\frac{\partial\psi_6(W_i,\hat{\theta}^w)}{\partial\beta_1'} & \frac{\partial\psi_6(W_i,\hat{\theta}^w)}{\partial\beta_0'} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{-1} & \mathbf{0} \\
\mathbf{0} & \mathbf{0} & \frac{\partial\psi_7(W_i,\hat{\theta}^w)}{\partial\alpha_1'} & \frac{\partial\psi_7(W_i,\hat{\theta}^w)}{\partial\alpha_0'} & \mathbf{0} & \mathbf{0} & \mathbf{-1}
\end{pmatrix}
\end{aligned}$$

For the linear mean function $m_z(X_i, \beta_z)$ and the logit specifications for the mean function of the treatment indicator $\mu_z(X_i; \alpha_z)$ and the instrument propensity score

$P(X_i;\gamma)$ the derivatives of the moment functions are:

$$
\begin{aligned}
\frac{\partial\psi_1(W_i,\hat{\theta}^w)}{\partial\beta_1'} &= \frac{Z_i}{\Lambda(X_i'\hat{\gamma}^w)}X_iX_i' \\
\frac{\partial\psi_1(W_i,\hat{\theta}^w)}{\partial\gamma'} &= -\frac{Z_i}{\Lambda(X_i'\hat{\gamma}^w)^2}\Lambda(X_i'\hat{\gamma}^w)(1-\Lambda(X_i'\hat{\gamma}^w))(Y_i-X_i'\hat{\beta}_1^w)X_iX_i' \\
\frac{\partial\psi_2(W_i,\hat{\theta}^w)}{\partial\beta_0'} &= \frac{1-Z_i}{1-\Lambda(X_i'\hat{\gamma}^w)}X_iX_i' \\
\frac{\partial\psi_2(W_i,\hat{\theta}^w)}{\partial\gamma'} &= \frac{1-Z_i}{(1-\Lambda(X_i'\hat{\gamma}^w))^2}\Lambda(X_i'\hat{\gamma}^w)(1-\Lambda(X_i'\hat{\gamma}^w))(Y_i-X_i'\hat{\beta}_0^w)X_iX_i' \\
\frac{\partial\psi_3(W_i,\hat{\theta}^w)}{\partial\alpha_1'} &= \frac{Z_i}{\Lambda(X_i'\hat{\gamma}^w)}\Lambda(X_i'\hat{\alpha}_1^w)(1-\Lambda(X_i'\hat{\alpha}_1^w))X_iX_i' \\
\frac{\partial\psi_3(W_i,\hat{\theta}^w)}{\partial\gamma'} &= -\frac{Z_i}{\Lambda(X_i'\hat{\gamma}^w)^2}\Lambda(X_i'\hat{\gamma}^w)(1-\Lambda(X_i'\hat{\gamma}^w))(D_i-\Lambda(X_i'\hat{\alpha}_1^w))X_iX_i' \\
\frac{\partial\psi_4(W_i,\hat{\theta}^w)}{\partial\alpha_0'} &= \frac{1-Z_i}{1-\Lambda(X_i'\hat{\gamma}^w)}\Lambda(X_i'\hat{\alpha}_0^w)(1-\Lambda(X_i'\hat{\alpha}_0^w))X_iX_i' \\
\frac{\partial\psi_4(W_i,\hat{\theta}^w)}{\partial\gamma} &= \frac{1-Z_i}{(1-\Lambda(X_i'\hat{\gamma}^w))^2}\Lambda(X_i'\hat{\gamma}^w)(1-\Lambda(X_i'\hat{\gamma}^w))(D_i-\Lambda(X_i'\hat{\alpha}_0^w))X_iX_i' \\
\frac{\partial\psi_5(W_i,\hat{\theta}^w)}{\partial\gamma} &= -\Lambda(X_i'\hat{\gamma}^w)(1-\Lambda(X_i'\hat{\gamma}^w))X_iX_i' \\
\frac{\partial\psi_6(W_i,\hat{\theta}^w)}{\partial\beta_1'} &= X_i' \\
\frac{\partial\psi_6(W_i,\hat{\theta}^w)}{\partial\beta_0'} &= -X_i' \\
\frac{\partial\psi_7(W_i,\hat{\theta}^w)}{\partial\alpha_1'} &= \Lambda(X_i'\hat{\alpha}_1^w)(1-\Lambda(X_i'\hat{\alpha}_1^w))X_i' \\
\frac{\partial\psi_7(W_i,\hat{\theta}^w)}{\partial\alpha_0'} &= -\Lambda(X_i'\hat{\alpha}_0^w)(1-\Lambda(X_i'\hat{\alpha}_0^w))X_i'
\end{aligned}
$$

B Tables

Table B1: Simulation Results

	n=500			n=1000			n=4000		
	IPS	REG	DR	IPS	REG	DR	IPS	REG	DR
				A: All specifications are correct					
BIAS	*-0.01*	*-0.08*	*-0.02*	*-0.01*	*-0.07*	*-0.01*	*0.00*	*-0.06*	*0.00*
MRBIAS	*0.61*	*0.44*	*0.49*	*0.38*	*0.30*	*0.34*	*0.18*	*0.15*	*0.16*
SE	*1.04*	*0.81*	*0.89*	*0.69*	*0.55*	*0.61*	*0.34*	*0.27*	*0.30*
MSE	*1.17*	*0.64*	*0.81*	*0.50*	*0.32*	*0.39*	*0.11*	*0.08*	*0.09*
				B: Correct specification of $E[Y\|X,Z]$ and $E[D\|X,Z]$					
				Wrong specification of Instrument Propensity Score $P(Z=1\|X)$					
BIAS	1.97	*-0.08*	*-0.09*	2.38	*-0.07*	*-0.07*	2.25	*-0.06*	*-0.07*
MRBIAS	2.45	*0.44*	*0.45*	1.73	*0.30*	*0.31*	1.61	*0.15*	*0.15*
SE	138.92	*0.81*	*0.83*	1.25	*0.55*	*0.57*	0.57	*0.27*	*0.28*
MSE	1637.35	*0.64*	*0.68*	7.30	*0.32*	*0.34*	5.39	*0.08*	*0.08*
				C: Correct specification of Instrument Propensity Score $P(Z=1\|X)$					
				C1: Wrong specification of $E[Y\|X,Z]$ and $E[D\|X,Z]$					
BIAS	*-0.02*	2.71	*-0.01*	*-0.01*	2.41	*0.00*	*0.00*	2.27	*0.01*
MRBIAS	*0.57*	2.05	*0.49*	*0.38*	1.74	*0.34*	*0.18*	1.63	*0.16*
SE	*1.03*	3.28	*0.90*	*0.69*	1.20	*0.62*	*0.34*	0.54	*0.30*
MSE	*1.08*	25.50	*0.82*	*0.50*	7.34	*0.40*	*0.11*	5.44	*0.09*
				C2: Wrong specification of $E[D=1\|X,Z]$, correct specification of $E[Y\|X,Z]$					
BIAS	*-0.02*	0.70	*-0.02*	*-0.01*	0.63	*-0.01*	*0.00*	0.60	*0.00*
MRBIAS	*0.57*	0.86	*0.49*	*0.38*	0.60	*0.33*	*0.18*	0.44	*0.16*
SE	*1.03*	1.95	*0.89*	*0.69*	0.90	*0.61*	*0.34*	0.42	*0.30*
MSE	*1.08*	5.14	*0.80*	*0.50*	1.25	*0.39*	*0.11*	0.53	*0.09*
				C:3 Wrong specification of $E[Y\|X,Z]$, correct specification of $E[D=1\|X,Z]$					
BIAS	*-0.02*	1.09	*-0.01*	*-0.01*	1.07	*0.00*	*0.00*	1.05	*0.01*
MRBIAS	*0.57*	0.87	*0.50*	*0.38*	0.79	*0.34*	*0.18*	0.76	*0.16*
SE	*1.03*	1.04	*0.91*	*0.69*	0.70	*0.62*	*0.34*	0.34	*0.30*
MSE	*1.08*	1.98	*0.83*	*0.50*	1.52	*0.40*	*0.11*	1.19	*0.09*

Note: The results based on 5000 Monte Carlo Samples. X's are uncorrelated

Table B2: Simulation Results

	n=500			n=1000			n=4000		
	IPS	REG	DR	IPS	REG	DR	IPS	REG	DR
				A: All specifications are correct					
BIAS	*-0.01*	*-0.04*	*-0.01*	*-0.01*	*-0.03*	*-0.01*	*0.00*	*-0.02*	*0.00*
MRBIAS	*0.45*	*0.41*	*0.43*	*0.30*	*0.28*	*0.29*	*0.15*	*0.14*	*0.14*
SE	*0.84*	*0.75*	*0.79*	*0.56*	*0.51*	*0.54*	*0.27*	*0.25*	*0.26*
MSE	*0.67*	*0.55*	*0.61*	*0.33*	*0.27*	*0.30*	*0.07*	*0.06*	*0.07*
				B: Correct specification of $E[Y\|X,Z]$ and $E[D\|X,Z]$					
				Wrong specification of Instrument Propensity Score $P(Z=1\|X)$					
BIAS	*1.74*	*-0.04*	*-0.06*	1.87	*-0.03*	*-0.06*	1.79	*-0.02*	*-0.05*
MRBIAS	1.76	*0.41*	*0.41*	1.37	*0.28*	*0.28*	1.30	*0.14*	*0.14*
SE	43.56	*0.75*	*0.75*	0.99	*0.51*	*0.51*	0.45	*0.25*	*0.25*
MSE	478.18	*0.55*	*0.55*	4.53	*0.27*	*0.28*	3.40	*0.06*	*0.06*
				C: Correct specification of Instrument Propensity Score $P(Z=1\|X)$					
				C1: Wrong specification of $E[Y\|X,Z]$ and $E[D\|X,Z]$					
BIAS	*-0.01*	2.00	*0.00*	*-0.01*	1.88	*0.00*	*0.00*	1.79	*0.00*
MRBIAS	*0.45*	1.58	*0.43*	*0.30*	1.37	*0.29*	*0.15*	1.30	*0.14*
SE	*0.84*	2.60	*0.79*	*0.56*	0.97	*0.54*	*0.27*	0.45	*0.26*
MSE	*0.67*	19.57	*0.62*	*0.33*	4.55	*0.30*	*0.07*	3.41	*0.07*
				C2: Wrong specification of $E[D=1\|X,Z]$, correct specification of $E[Y\|X,Z]$					
BIAS	*-0.01*	0.55	*-0.01*	*-0.01*	0.52	*-0.01*	*0.00*	0.51	*0.00*
MRBIAS	*0.45*	0.73	*0.43*	*0.30*	0.51	*0.29*	*0.15*	0.38	*0.14*
SE	*0.84*	1.74	*0.78*	*0.56*	0.77	*0.54*	*0.27*	0.36	*0.26*
MSE	*0.67*	5.17	*0.61*	*0.33*	0.91	*0.30*	*0.07*	0.39	*0.07*
				C3: Wrong specification of $E[Y\|X,Z]$, correct specification of $E[D=1\|X,Z]$					
BIAS	*-0.01*	0.93	*0.00*	*-0.01*	0.91	*0.00*	*0.00*	0.90	*0.00*
MRBIAS	*0.45*	0.75	*0.43*	*0.30*	0.68	*0.29*	*0.15*	0.65	*0.14*
SE	*0.84*	0.97	*0.80*	*0.56*	0.65	*0.54*	*0.27*	0.31	*0.27*
MSE	*0.67*	1.55	*0.62*	*0.33*	1.15	*0.31*	*0.07*	0.88	*0.07*

Note: The results based on 5000 Monte Carlo Samples. X's are correlated

Table B3: Variable Descriptions

Label	Description
Outcome	
LNWAGE	log hourly wages
Treatment	
D	= 1 if the individual has a upper secondary school diploma, = 0 otherwise
Instrument	
Z	= 1 if the individual did ***not*** repeat the 10th grade = 0 otherwise
Covariates	
FEMALE	= 1 if female, = 0 otherwise
AGE	Age in years
IQ	Number of correctly solved questions in the Intelligence Structure Test (IST; Amthauer (1953)). The test was carried out in 1969.
PR_RET	= 1 if the individual did repeat a grade prior to the 10th grade = 0 otherwise
AGEM	Age of the mother in 1970
DILIG	Measure of attributing success to diligence on a scale from 0 (weaker) to 5 (stronger)
LUCK	Measure of attributing success to luck on a scale from 0 (weaker) to 5 (stronger)
FAMILY	Measure of attributing success to the family on a scale from 0 (weaker) to 5 (stronger)
ABIL	Measure of attributing success to ability on a scale from 0 (weaker) to 5 (stronger)
ASTUTE	Measure of attributing success to astuteness on a scale from 0 (weaker) to 5 (stronger)
EDU_MOT	Categorical variable for educational attainment of the mother from 1-4
EDU_FAT	Categorical variable for educational attainment of the father from 1-4
HHINC	Categorical variable for net household income in 1970 from 1-9 =1 up to 750 DM, =2 751 up to 1000 DM, =3 1001 up to 1250 DM, =4 1251 up to 1500 DM, =5 1501 up to 2000 DM,=6 2001 up to 2500 DM, =7 2501 up to 3000 DM, =8 3001 up to 4000 DM,=9 more than 4000 DM
EMP_MOT	Categorical variable for mother's employment status from 1-3
PARINT1	Dummy, 1 if parents are interested in promotion on to the next grade level
PARINT2	Dummy, 1 if parents are interested in final grades
PARINT3	Dummy, 1 if parents are interested in test grades
INTSCHOOL	Average value of PARINT1, PARINT2 and PARINT3
AGEMOT	Categorical variable for mother's age from 1-9 =1 if 30-34, =2 if 35-39, =3 if 40-44, =4 if 45-49, =5 if 50-54, =6 if 55-59, =7 if 60-64, =8 if 65-70, =9 if she died
WISH	Do you want to continue studying after upper secondary school? =1 if the answer is yes, =2 if maybe, =3 if no, =4 if do not know yet, =5 if no upper secondary school degree is planned

Table B4: Descriptive Statistics

	Entire Sample	By Graduation		By Grade Retention	
		D=1	D=0	Z=1	Z=0
Outcome					
LNWAGE	2.93	3.00	2.78	2.94	2.82
	(0.43)	(0.44)	(0.39)	(0.44)	(0.34)
Treatment					
D	0.67	1.00	0.00	0.71	0.30
	(0.47)			(0.45)	(0.46)
Instrument					
Z	0.88	0.95	0.75	1.00	0.00
	(0.32)	(0.22)	(0.43)		
Covariates					
EMP_MOT	2.01	2.03	1.98	2.02	1.92
	(0.70)	(0.69)	(0.72)	(0.69)	(0.73)
AGEM	3.61	3.64	3.56	3.64	3.40
	(1.18)	(1.18)	(1.20)	(1.17)	(1.26)
AGE	15.40	15.20	15.81	15.39	15.47
	(0.89)	(0.83)	(0.88)	(0.90)	(0.86)
IQ	41.02	41.872	39.342	41.192	39.812
	(9.03)	(9.23)	(8.37)	(9.08)	(8.58)
EDU_FAT	5.69	6.16	4.75	5.71	5.50
	(4.22)	(4.35)	(3.79)	(4.24)	(4.07)
EDU_MOT	4.08	4.42	3.41	4.10	3.99
	(3.47)	(3.65)	(2.95)	(3.50)	(3.17)
PR_RET	0.35	0.24	0.55	0.34	0.40
	(0.48)	(0.43)	(0.50)	(0.47)	(0.49)
INTSCHOOL	0.66	0.68	0.62	0.66	0.62
	(0.30)	(0.30)	(0.29)	(0.30)	(0.29)
WISH	2.65	2.15	3.63	2.60	2.98
	(1.58)	(1.33)	(1.57)	(1.56)	(1.63)
DILIG	4.11	4.07	4.20	4.12	4.08
	(1.04)	(1.07)	(0.99)	(1.02)	(1.22)
LUCK	2.21	2.16	2.32	2.19	2.36
	(1.59)	(1.56)	(1.64)	(1.58)	(1.68)
ABIL	3.52	3.53	3.51	3.54	3.42
	(1.07)	(1.09)	(1.04)	(1.06)	(1.16)
ASTUTE	3.11	3.11	3.10	3.08	3.33
	(1.31)	(1.31)	(1.32)	(1.31)	(1.32)
FAMILY	2.08	2.01	2.21	2.04	2.32
	(1.54)	(1.52)	(1.58)	(1.53)	(1.62)
HHINC	4.40	4.56	4.07	4.39	4.43
	(2.07)	(2.09)	(1.97)	(2.08)	(1.97)
Number of obs.	1552	1033	519	1369	183

Note: Standard errors are given in parentheses

Table B5: Logit Estimation Results

Dependent Variable Z			
Variables	Coef.	Std. Err.	p-value
FEM	0.418	0.17	0.02
EMP_MOT	0.213	0.12	0.07
AGEM	0.195	0.07	0.01
AGE	0.041	0.11	0.71
IQ	0.018	0.01	0.05
EDU_VAT	-0.004	0.03	0.89
EDU_MOT	0.004	0.03	0.91
PR_RET	-0.175	0.20	0.39
INTSCHOOL	0.383	0.28	0.16
WISH	-0.169	0.06	0.00
DILIG	0.088	0.08	0.25
LUCK	-0.009	0.05	0.87
ABIL	0.124	0.07	0.10
FAMILY	-0.125	0.05	0.02
ASTUTE	-0.159	0.07	0.02
HHINC	-0.028	0.05	0.58
CONS	-0.203	1.83	0.91
n = 1552			
Wald $\chi^2(16) = 49.55$			
Prob $> \chi^2 = 0.0000$			
Log pseudolikelihood = -539.55216			
Pseudo $R^2 = 0.0416$			

C Figures

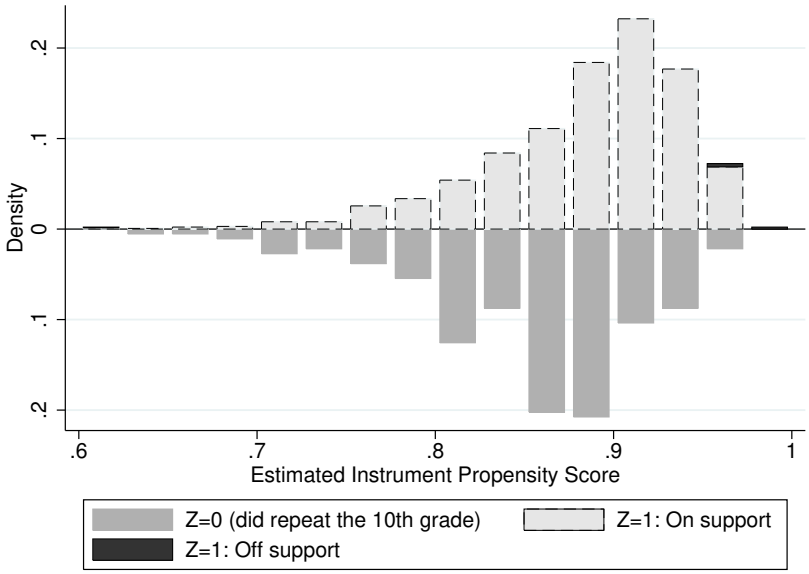

Figure C1: Density of Estimated Probability of Grade Retention. Estimation is based on specification given in Table B 5