Uysal, Selver Derya

Conference Paper

The Effect of Grade Retention on School Outcomes: An Application of Doubly Robust Estimation Methods

Beiträge zur Jahrestagung des Vereins für Socialpolitik 2010: Ökonomie der Familie - Session: Evaluation Econometrics, No. A6-V3

Provided in Cooperation with:
Verein für Socialpolitik / German Economic Association

Suggested Citation: Uysal, Selver Derya (2010) : The Effect of Grade Retention on School Outcomes: An Application of Doubly Robust Estimation Methods, Beiträge zur Jahrestagung des Vereins für Socialpolitik 2010: Ökonomie der Familie - Session: Evaluation Econometrics, No. A6-V3, Verein für Socialpolitik, Frankfurt a. M.

This Version is available at:
https://hdl.handle.net/10419/37510

The Effect of Grade Retention on School Outcomes: An Application of Doubly Robust Estimation Methods

Selver Derya Uysal*

Department of Economics

University of Konstanz

August 27, 2010
THIS VERSION for "VfS Jahrestagung".
PLEASE DO NOT QUOTE.

Abstract

In the first part of this study, we estimate the average causal treatment effect of grade retention on several educational outcome variables, such as completion of upper secondary school, graduation grades in math and German, as well as average final grade using a data set from Germany. We use doubly robust method, regression adjustment and inverse propensity score weighting. The results of the empirical study show that grade retention does not improve the students' educational achievement. Furthermore, in the second part this study 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 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 10^{th} grade.
JEL classification: $C21, I2$

Keywords: Econometric evaluation, doubly robust estimation, CIA, ATE, LATE, propensity score, grade retention

*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).

1 Introduction

Grade retention is an intervention tool in education. It refers to the practice of requiring a student to repeat the same grade which s/he has already completed because of her/his poor performance. In Jackson (1975), the aim of grade retention is explained as an attempt at remedying inadequate academic progress and contributing to the development of students not ready for the next grade. The underlying idea is that students who do not successfully complete a grade level will not be able to digest the next higher grade's material. These students are therefore, for their own interest, required to repeat the grade. The most important question, however, is whether grade retention really helps students to improve their grades or whether it harms the students' school success. This paper aims to address this question and estimates the causal effect of this school intervention on several school outcomes.

The effects of grade retention have been a discussion topic for more than four decades. Most studies concentrate on the effects of grade retention on performance in later grades, on the likelihood of drooping out of high school, and on labor market outcomes for late adolescence (see Guevremont, Roos, and Brownell (2007), McCoy and Reynolds (1999), Jimerson (1999), Jimerson (2001), and Eide and Showalter (2001) among others.). The results are somewhat controversial: although the vast majority of empirical work done with the data from the US and Canada points out the negative effects of grade retention, there are also a number of papers indicating gains.

Since being held in a grade is not a random assignment, simple mean comparisons of outcome variables do not reveal the true causal effect of grade retention. We could realize true causal effects over a whole population by using mean comparisons, if we could randomly hold schoolchildren in the same grade for a second year. Since such an experiment on schoolchildren is impossible and unethical, we should rely on the econometric methods which enable identification of the true causal effects in terms of potential outcomes. In the case of binary treatment, there are two potential outcomes for treated and nontreated cases: one observed depending on the realized treatment status, and the other one unobserved (i.e. counterfactual). Identification is achieved under some assumptions in potential outcome framework. The crucial assumption we use in this paper is Conditional Independence Assump-

tion[1]. It means given a set of observable characteristics which are not affected by the treatment, potential outcomes are independent of treatment assignment. There are several methods proposed for estimating treatment effects under the assumption of conditional independence (see Imbens (2004) for a review). The main methods can be categorized into regression, propensity score weighting and matching methods. Here, we estimate the effect of grade retention on different outcomes using regression, propensity score weighting and a combination of regression and propensity score weighting methods. The advantage of the combination over the single methods is that the mixed method provides double protection against misspecification. That is, the estimator is still consistent, even if either the propensity score or the mean function is wrongly specified but not both (for further discussion of double robustness see Robins, Rotnitzky, and Zhao (1995), Robins and Ritov (1997), Hirano and Imbens (2001), Wooldridge (2007), and Bang and Robins (2005)).

In the second part of this paper, we extend the use of the doubly robustnees property in the estimation of Local Average Treatment Effect. 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.[2] 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 second part, we propose a doubly robust estimation method of the LATE with covariates. The identifying assumptions are similar to those of Abadie (2003), Tan (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).

[1]This assumption is called *Ignorability of Treatment* (given observed covariates X) by Rosenbaum and Rubin (1983) and *Unconfoundedness* by Imbens (2004).

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

In this paper we use a German dataset "Gymnasiastenstudie" (Central Archive for Empirical Social Research (2007)) in order to estimate the causal effect of grade retention on different school outcomes. This work distinguishes from the existing literature in many ways. First of all, to our knowledge, there is no empirical study published which analyzes the effects of grade retention using a Germany dataset. The dataset we use here is restricted to students attending upper secondary school (Gymnasium) in North Rhine-Westphalia. However, it is still representative for Germany, since one fourth 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 that the upper secondary schools (Gymnasien) in Germany serve almost for one half of the total students after primary education (Grundschule)[3]. This paper is also one of the very few papers which rely on econometric evaluation methods in order to analyze the effect of grade retention on school outcomes. Another contribution of this paper is that it uses one of the least applied econometric evaluation methods, namely Doubly Robust Method[4].

The organization of the paper is as follows: Section 2 briefly explains identifying assumptions and the econometric methods applied. In this section, we explain existing doubly robust methods for estimation of the ATE and propose our extension of this kind of methods for estimation of the LATE. Section 3.1 focuses on the sample and elaborates on the empirical results. This section has also two parts: estimation of the ATE and estimation of the LATE with the proposed method. Finally, Section 4 summarizes the main results and concludes the paper.

2 Econometric Method

2.1 Estimation of the ATE

Consider N units which are drawn from a large population. For each individual i in the sample, where $i = 1, ..., N$, we observe the triple (Y_i, D_i, X_i). D_i shows the

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

[4]To our knowledge, there are only two applications of this method: Bang and Robins (2005) and Hirano and Imbens (2001)

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 observe also a vector of characteristics (covariates) for the i^{th} individual denoted by X_i. For each individual there are two potential outcomes (Y_{i0}, Y_{i1}). Y_{id} denotes the outcome for each individual i, for which $D_i = d$ where $d \in \{0, 1\}$. For each individual only one of the potential outcomes is observed depending on the treatment status. The observed outcome, denoted by Y_i in the triple, can be written in terms of treatment indicator (D_i) and the potential outcomes:

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

Our primary interest lies in estimating the average causal effect of the repeating a grade. This effect is called the average treatment effect (ATE). It gives the mean effect of the treatment:

$$\tau = E[Y_{i1} - Y_{i0}] = E[Y_{i1}] - E[Y_{i0}]$$

Since only one of the potential outcomes is observed, ATE cannot be identified without further assumptions. For the empirical study we assume that the following assumptions hold:

A 2.1 *Conditional Independence Assumption (CIA)*
$Y_{i0}, Y_{i1} \perp D_i | X_i$, *where* $\perp$ *stands for independence.*

It implies that after controlling for the effect of covariates, treatment and outcomes are independent.

A 2.2 *Common Support*
$0 < Pr(D_i = 1 | X_i) < 1$

Assumption 2.2 means that for all x there is a positive probability of either participating $(D_i = 1)$ or not participating $(D_i = 0)$. In other words for each value of covariates there are both treated and untreated cases. Thus, there is an overlap between the treated and untreated subsamples. If the assumption fails, then we could have individuals with x vectors who are all treated and those with a different

x vector who are all untreated.

Rosenbaum and Rubin (1983) show that under CIA identification can be achieved by conditioning on a function of X_i, a balancing score[5], instead of a high dimensional X_i itself. The most commonly used balancing score in the evaluation literature is the propensity score, the conditional probability of assignment to the treatment given the covariates:

$$p(x) = Pr[D_i = 1|X_i = x] = E[D_i|X_i = x] \tag{2.1}$$

Lemma 2.1 *Unconfoundedness Given the Propensity Score*
Given the CIA and Common Support assumptions, outcomes Y_{i0} and Y_{i1} are independent of treatment given the propensity score.
$Y_{i0}, Y_{i1} \perp D_i|p(X_i)$

Under these assumptions several methods can be used to estimate the average treatment effect. This paper uses three different methods: regression method, inverse propensity score weighting method and Doubly Robust Method which is the combination of the first two methods.

Under the CIA one can estimate the unconditional means $E[Y_{id}] = \mu_d$ based on the parametric estimation of conditional means $E[Y_{id}|X_i = x]$ for $d \in 0, 1$. Since the arguments are symmetric, we concentrate on $E[Y_{i1}|X_i = x]$. Assume that the conditional mean function is correctly specified, $E[Y_{i1}|X_i = x] = m_1(x, \beta_1)$, where $m_1(x, \beta_1)$ is a function depending on a covariate vector and a $k-$dimensional true parameter vector β_1. Given a consistent estimator $\hat{\beta}_1$, a consistent estimator of the unconditional mean, μ_1, is:

$$\hat{\mu}_1 = \frac{1}{N} \sum_i m_1(X_i, \hat{\beta}_1) \tag{2.2}$$

since $\mu_1 = E[m_1(x, \beta_1)]$ by iterated expectations.

[5]A balancing score is a function of observed covariates X_i such that the conditional distribution of X_i given balancing score is the same for treated and control units (see Rosenbaum and Rubin (1983)).

Thus, one can estimate the average treatment effect based on two parametric regressions as follows:

$$\hat{\tau}_{reg} = \frac{1}{N}\sum_i [m_1(X_i, \hat{\beta}_1) - m_0(X_i, \hat{\beta}_0)] \tag{2.3}$$

From Wooldridge (2002) and Wooldridge (2009), the asymptotic variance can be written as follows:

$$\begin{aligned} AV\sqrt{N}(\hat{\tau}_{reg}) = E[(m_1(X,\beta_1) - m_0(X,\beta_0) - \tau_{reg})^2] \\ + E[\frac{\partial m_1(X,\beta_1)}{\partial \beta_1'}]V_1 E[\frac{\partial m_1(X,\beta_1)}{\partial \beta_1'}]' \\ + E[\frac{\partial m_0(X,\beta_0)}{\partial \beta_0'}]V_0 E[\frac{\partial m_0(X,\beta_0)}{\partial \beta_0'}]' \end{aligned} \tag{2.4}$$

where V_1 and V_0 are the variances of β_1 and β_0. The variance can be estimated by replacing the expectations with the sample means and true parameters with their estimates.

Using Lemma 2.1, the mean outcomes for the treatment and control groups can be identified by weighting the observations with the inverse of the propensity score:

$$E[Y_{i1}] = E[DY/p(X)]$$

$$E[Y_{i0}] = E[(1-D)Y/(1-p(X))]$$

Hence, we can write the ATE as follows:

$$\tau = E[\frac{DY}{p(X)} - \frac{(1-D)Y}{(1-p(X))}]$$

The estimator of ATE can be written as a sample counterpart of the population expectation. Usually this estimator is referred as the propensity score weighting estimator[6]:

$$\hat{\tau}_{ps} = \frac{1}{n}\sum_{i=1}[D_i Y_i/p(X_i;\hat{\alpha}) - (1-D_i)Y_i/(1-p(X_i;\hat{\alpha}))] \tag{2.5}$$

$$= \frac{1}{n}\sum_{i=1}\frac{(D_i - p(X_i;\hat{\alpha}))Y_i}{p(X_i;\hat{\alpha})(1-p(X_i;\hat{\alpha}))} \equiv \frac{1}{n}\sum_{i=1}\hat{g}_i \tag{2.6}$$

[6]This estimator is identical to an estimator from Horvitz and Thompson (1952) for handling nonrandom sampling.

Since usually the true propensity score $p(X)$ is not observable, one can use an estimated propensity score $p(X;\hat{\alpha})$, where $\hat{\alpha}$ is the maximum likelihood estimator (MLE) (e.g., probit or logit) of the parameter vector of the propensity score specification. $\hat{\tau}_{ps}$ is inconsistent, however, if the propensity score is misspecified (see for further discussion Horvitz and Thompson (1952), Rosenbaum (1987), and Bang and Robins (2005))[7].

Following Wooldridge (2007), Wooldridge (2009) shows that the asymptotic variance of τ_{ps} is:

$$AV\sqrt{N}(\hat{\tau}_{ps} - \tau) = E[e_i e_i'] \tag{2.7}$$

where $e_i \equiv g_i - E[g_i s_i']E[s_i s_i']^{-1} s_i$, s_i is the score function of the MLE model of the propensity score.

Both of the above mentioned estimation methods, regression and propensity score weighting, can be easily implemented. There are no computational difficulties, or curse of dimensionality problems as in nonparametric methods. As mentioned above, consistency of the estimates hinges upon the true specification of the mean or the propensity score, depending on which estimation method is used. Wooldridge (2007) and Hirano and Imbens (2001) show, however, that combining weighting and regression methods gives a doubly robust estimate of the unconditional mean, providing double protection against misspecification. As long as one of the functional form specifications, either that for the conditional mean or the propensity score, is correctly specified, the resulting estimator for the unconditional mean will be consistent provided that $E[Y_d] = E[m_d(x, \beta_d^*)]$ where β_d^* is the probability limit of an estimator from the conditional mean function (Wooldridge (2007)). This property holds for linear exponential family with a canonical link function (see for details Wooldridge (2007), Scharfstein, Rotnitzky, and Robins (1999)). The three regression models we use for this application, namely linear, logit and poisson regression, belong to this family.

The main idea is weighting the objective function of the regression by the inverse of the propensity score. Depending on the choice of the regression method, the coefficient estimates of the mean function parameters come from weighted least square or weighted MLE method. The score function of the chosen parametric model is

[7]Hirano, Imbens, and Ridder (2003) examine the estimator in equation 2.5 where $p(X_i;\hat{\alpha})$ is replaced by nonparametric estimates.

weighted by $1/p(X_i;\hat{\alpha})$ and by $1/(1-p(X_i;\hat{\alpha}))$ for treated and untreated subpopulation respectively.

Depending on the nature of outcome variable the proper mean function is one of the following:

- For a continuous outcome variable:

$$m_d(X_i, \beta_{dw}) = X_i'\beta_{dw} \tag{2.8}$$

- For a binary outcome variable:

$$m_d(X_i, \beta_{dw}) = \Lambda(X_i'\beta_{dw}) = \frac{\exp(X_i'\beta_{dw})}{1+\exp(X_i'\beta_{dw})} \tag{2.9}$$

- For a count outcome variable:

$$m_d(X_i, \beta_{dw}) = \exp(X_i'\beta_{dw}) \tag{2.10}$$

The estimated coefficient $\hat{\beta}_{dw}$ from weighted regression method solves the weighted score function

$$\frac{1}{N}\sum_i w_i(Y_i - m_d(X_i, \hat{\beta}_{dw}))X_i = 0 \tag{2.11}$$

where

$$w_i = \begin{cases} 1/p(X_i;\hat{\alpha}), & \text{if} \quad D_i = 1 \\ 1/(1-p(X_i;\hat{\alpha})), & \text{if} \quad D_i = 0 \end{cases}$$

Thus, one can estimate the average treatment effect based on two weighted regression coefficients as in regression methods:

$$\hat{\tau}_{dr} = \frac{1}{N}\sum_i [m_1(X_i, \hat{\beta}_{1w}) - m_0(X_i, \hat{\beta}_{0w})] \tag{2.12}$$

The asymptotic variance of $\hat{\tau}_{dr}$ is same as Equation 2.4 with different V_0 and V_1[8]. When estimating V_0 and V_1, one has to take into account that the weights are

[8]For the linear case the asymptotic variance of $\hat{\tau}_{dr}$ is equivalent to the variance derived by Hirano and Imbens (2001) for linear mean function.

estimated in a first step. Wooldridge (2007) derives the asymptotic variance of $\hat{\beta}_{dw}$ as follows.

$$AV\sqrt{N}(\hat{\beta}_{dw}) = A_0^{-1} D_0 A_0^{-1}$$

where $A_0 \equiv E[H(X, \beta_{dw})]$ and $D_0 \equiv E[k_i k_i']$. $k_i = k(X_i, \beta_{dw}) = w_i(Y_i - m(X_i, \hat{\beta}_{dw}))X_i$ is the weighted score function and $H(X, \beta_{dw})$ is the Hessian. Wooldridge (2007) proposes also the following consistent estimators for A_0 and D_0 :

$$\hat{A} = \frac{1}{N} \sum_i w_i H(X_i, \hat{\beta}_{dw})$$

$$\hat{D} = \frac{1}{N} \sum_i k(X_i, \hat{\beta}_{dw}) k(X_i, \hat{\beta}_{dw})'$$

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

$$Y_i = D_i Y_{i1} + (1 - D_i) Y_{i0} \tag{2.13}$$

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:

$$D_i = Z_i D_i^1 + (1 - Z_i) D_i^0. \tag{2.14}$$

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:

$$\tau_{LATE} \quad = \quad E[Y_1 - Y_0 | D^1 > D^0]. \tag{2.15}$$

In order to identify the LATE we use the following assumptions:[9]

A 2.3 *Conditional Independence of the Instrument:*
$(Y_0, Y_1, D^z) \perp Z|X$ *for each* $z \in \{0, 1\}$.

A 2.4 *Rank Condition:*
$\Pr[D = 1 | X, Z]$ *is a nontrivial function of* Z*, conditional on* $X = x$.

A 2.5 *Monotonicity:*
$\Pr[D^1 \geq D^0 | X] = 1$.

A 2.6 *First Stage:*
$0 < \Pr[Z = 1 | X = x] < 1$ *and* $\Pr[D^1 = 1 | X] > \Pr[D^0 = 1 | X]$.

Assumption A 2.3 implies that the instrument, Z, is as good as randomly assigned once we condition on the covariates, X. Assumption A 2.3 also rules out the direct effect of the instrument on the potential outcome. The first two assumption

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

together guarantee that the only effect of the instrument on the outcome is through the treatment variable. Assumption A 2.6 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 2.5 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.

Imbens and Angrist (1994) show that if Assumptions A 2.3-2.5 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:

$$\tau_{LATE} = \frac{E[Y|Z=1] - E[Y|Z=0]}{E[D|Z=1] - E[D|Z=0]}. \tag{2.16}$$

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

Theorem 1 *Under Assumptions A 2.3-A 2.5 the conditional LATE is identified as*

$$\tau_{LATE}(x) = \mathrm{E}\left[Y_1 - Y_0 | \, X = x, D^1 > D^0\right] = \frac{E[Y|X=x, Z=1] - E[Y|X=x, Z=0]}{E[D|X=x, Z=1] - E[D|X=x, Z=0]} \tag{2.17}$$

Theorem 2 *Under Assumptions A 2.3-A 2.6 the unconditional LATE is identified as follows:*

$$\tau_{LATE} = \frac{\mathrm{E}_X\left[\mathrm{E}\left[Y | \, X, Z=1\right] - \mathrm{E}\left[Y | \, X, Z=0\right]\right]}{\mathrm{E}_X\left[\mathrm{E}\left[D | \, X, Z=1\right] - \mathrm{E}\left[D | \, X, Z=0\right]\right]}. \tag{2.18}$$

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:*

$$\tau_{LATE}^{we} = \mathrm{E}\left[Y_1 - Y_0 | \, X = x, D^1 > D^0\right] = \frac{E\left[\frac{Z}{p(X)}Y\right] - E\left[\frac{1-Z}{1-p(X)}Y\right]}{E\left[\frac{Z}{p(X)}D\right] - E\left[\frac{1-Z}{1-p(X)}D\right]}. \tag{2.19}$$

These identification results show the connection between the LATE and the ATE. The ratio in Equation (2.17) 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 (2.17) and (2.18) 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. 2.13 and the conditional means of the outcome variable by substituting the definition of the potential treatment status as follows:

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

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 $\mathrm{E}[Y^z]$ and $\mathrm{E}[D^z]$ 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:

$$\hat{\tau}_{LATE}^{reg} = \frac{\frac{1}{N}\sum_i^N (m_1(X_i, \hat{\beta}_1) - m_0(X_i, \hat{\beta}_0))}{\frac{1}{N}\sum_i^N (\mu_1(X_i, \hat{\alpha}_1) - \mu_0(X_i, \hat{\alpha}_0))} = \frac{\hat{\tau}^{Y|Z}}{\hat{\tau}^{D|Z}}. \tag{2.20}$$

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. (2.19) 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:

$$\begin{aligned}\hat{\tau}^{we}_{LATE} &= \frac{N^{-1}\sum_i \left[\frac{Z_i}{\hat{p}(X_i;\hat{\gamma})}Y_i\right] - N^{-1}\sum_i \left[\frac{1-Z_i}{1-\hat{p}(X_i;\hat{\gamma})}Y_i\right]}{N^{-1}\sum_i \left[\frac{Z_i}{\hat{p}(X_i;\hat{\gamma})}D_i\right] - N^{-1}\sum_i \left[\frac{1-Z_i}{1-\hat{p}(X_i;\hat{\gamma})}D_i\right]} \\ &= \frac{N^{-1}\sum_i \frac{[Z_i-\hat{p}(X_i;\hat{\gamma})]Y_i}{\hat{p}(X_i;\hat{\gamma})[1-\hat{p}(X_i;\hat{\gamma})]}}{N^{-1}\sum_i \frac{[Z_i-\hat{p}(X_i;\hat{\gamma})]D_i}{\hat{p}(X_i;\hat{\gamma})[1-\hat{p}(X_i;\hat{\gamma})]}},\end{aligned} \tag{2.21}$$

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:

$$\sqrt{N}(\hat{\tau}_{LATE}-\tau_{LATE}) \xrightarrow{d} N\left(0, \left(\frac{1}{\tau^{D|Z}}\right)^2 V_{Y|Z} + \left(\frac{\tau^{Y|Z}}{(\tau^{D|Z})^2}\right)^2 V_{D|Z} - 2\frac{\tau^{Y|Z}}{(\tau^{D|Z})^3}\operatorname{Cov}\left(\tau^{Y|Z},\tau^{D|Z}\right)\right). \tag{2.22}$$

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.

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).[10] It is essentially the weighted regression of the four conditional means in Equation 2.20 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.

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

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

$$\hat{\tau}_{LATE}^{dr} = \frac{\frac{1}{N}\sum_i m_1(X_i, \hat{\beta}_1^w) - \frac{1}{N}\sum_i m_0(X_i, \hat{\beta}_0^w)}{\frac{1}{N}\sum_i \mu_1(X_i, \hat{\alpha}_1^w) - \frac{1}{N}\sum_i \mu_0(X_i, \hat{\alpha}_0^w)} \tag{2.23}$$

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,

[10] 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. 2.19 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 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. 2.19) or the conditional mean functions of the outcome variable and the treatment variable (Eq. 2.20). 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 2.19, 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:

$$\begin{aligned}
\mathrm{E}\left[\frac{Z_i}{p(X_i,\gamma)}(Y_i^1 - X_i'\beta_1)\right] &= \mathrm{E}\left[\mathrm{E}\left[\frac{Z_i}{p(X_i,\gamma)}(Y_i^1 - X_i'\beta_1)\middle| X_i\right]\right] \\
&= \mathrm{E}\left[\frac{\mathrm{E}\left[Z_i\middle| X_i\right]}{p(X_i,\gamma)}\mathrm{E}\left[(Y_i^1 - X_i'\beta_1)\middle| X_i\right]\right] \\
&= \mathrm{E}\left[\frac{\mathrm{E}\left[Z_i\middle| X_i\right]}{p(X_i,\gamma)}(\mathrm{E}\left[Y_i^1\middle| X_i\right] - X_i'\beta_1)\right].
\end{aligned} \tag{2.24}$$

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 $\mathrm{E}\left[Y_i^1\middle| X_i\right] = X_i'\beta_1$ holds, i.e., the expectation 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. (2.24) 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 $\text{Cov}\left(\tau^{Y|Z}, \tau^{Y|Z}\right)$, which we need in order to estimate the variance of the LATE estimator (see Eq. 2.22). 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

$$\frac{1}{N}\sum_{i=1}^{N}\psi(W_i, \hat{\theta}) = 0. \tag{2.25}$$

By standard results for M-Estimation it follows that:

$$\sqrt{N}(\hat{\theta} - \theta) \overset{a}{\sim} N\left(0, A^{-1}VA^{-1}\right) \tag{2.26}$$

where

$$A \equiv E\left[\frac{\partial\psi(W_i, \theta)}{\partial\theta'}\right] \tag{2.27}$$

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

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:

$$m_z(X_i, \beta_z) = g(X_i'\beta_z), \tag{2.28}$$

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. 2.25 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. (2.22). 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 Empirical Results

3.1 Grade Retention as a Treatment Variable

In the following, the causal effect of grade retention on several school outcomes is investigated for the German school system. The data set consists of information on family background and school related topics for about 3000 10^{th} grade students attending upper secondary school in North Rhine-Westphalia in the year 1970[11]. The students were sampled from 121 classes at 68 upper secondary schools. The data contains information from student, parent and teacher questionnaires. About ten years later, the students' grades were collected from the schools.

The empirical study on the causal effect of grade retention is distinguished from earlier studies by its investigation of the effect in a potential outcome framework and its application of the above explained methods for estimating the ATE of grade retention on school performance. Treatment is defined as repeating a class at least once after 10^{th} grade. The effects of grade retention on different outcome variables are investigated. The first one is the probability of graduating from upper secondary school (having "Abitur" or not). The other three outcome variables are only measured for those who have graduated from upper secondary school. One is the average final grade in upper secondary school. In addition, the effect on math and German "Abitur" grades is also considered. The aim of the empirical part is twofold: (i) estimate the causal effect of grade retention on the school performance, and (ii) investigate the differences of the causal effect for girls and boys. Outcomes are assumed to be independent of treatment status conditional on the covariates. All variables used in the study are listed in Table A1.

The variables are chosen in accordance with earlier findings concerning characteristics associated with being retained as well as with being successful in school. It is important to include variables related to both treatment status and potential outcomes so that the CIA holds approximately. A female dummy is included because most studies show that males are more likely to be retained than females. A measure of intelligence, IQ, is also included to control for the cognitive skills of the students. The variable IQ in our study is the sum of correctly solved questions of a standard psychometric Intelligence Structure Test (IST), which was adminis-

[11]The original data set consists of two more follow-ups in years 1984 and 1998.

tered in the class-room in the 10^th grade. Since noncognitive skills also appear to play an important role in school performance, as shown in earlier studies, variables which measure the attribution of success to diligence (DILIG) and ability (ABIL) are included as conditioning covariates. The variable WISH is added as a control for the child's motivation. We also control for the age of the student. Former studies also claim that the characteristics of parents, such as economic well being, education and parental involvement with their child's school performance, are also likely to affect the probability of being retained. EDU_MOT, EDU_FAT, AGE-MOT, HHINC, INTERSCHOOL are variables which control for family background and parents involvement. We can also identify whether the child has experienced any grade retention before 10^{th} grade (PR_RET).

The variables which are used in this study are chosen from three different sources. The outcome variables are taken from the administrative school data and the control variables are taken from parents and students questionnaires. Merging these three different data sets decreases the sample size already by about 500 observations. Some questions are asked to both students and parents. Thus, we combine the information sets to keep the decrease in the sample size moderate.

We create different samples. With the first sample we analyze the causal effect of grade retention on high school graduation (ABI) (see Table A2 for descriptive statistics). Thereafter, we restrict our sample for those who graduated from upper secondary school in order to estimate the causal effect of grade retention on average final grade (GPA) and final grades in math (MAT) and German (GER) (see Table A3 for descriptive statistics). Next we restrict the sample to the students who did not experience any grade retention before the 10^{th} grade (see Table A4) in order to see the effect of late grade retention on those students. For this sample we also look at the upper secondary school graduates and the effect of grade retention on graduation grades (see Table A5). For all four samples the analysis is done for the entire sample and for the subsamples by gender. The propensity score, the probability of being retained after 10^{th} grade is estimated by a logit regression for all subsamples. The regression results can be found in Table A6 and A7. Table A6 gives the logit estimation results for the sample before restricting by previous retention status and A7 gives the results only for students who did not experience retention before 10^{th} grade. From the logistic regression results, we can conclude that females are less likely to be retained. IQ has a decreasing effect on probability of being retained

in general. Having a young mother increases the probability of being retained at least for the main sample (Table A6 col. (a) and (c)). The variable PR_RET is highly significant and negative for the main sample (Table A6 col. (a), (b), (c)) . However, when we constrain our sample to high school graduates it does not have a significant effect on the probability of being hold in the same grade (Table A7 col. (a), (b), (c)). The variables, DILIG and ABIL, are also most of the time significantly negative. As in Rauber (2007), we also use these variables to measure to what extent a student follows an internal attribution strategy by attributing success to effort and ability. Relying on evidence that individuals with a high degree of self-esteem frequently tend to attribute success as being internal (see Rauber (2007) and its references), the interpretation of the negative coefficients might be that higher self esteem decreases the probability of grade retention. The other variable which is significantly negative for almost all samples is the willingness to pursue higher education (WISH), however with different signs for different subsamples. The coefficient (PARINT) which controls for parents interests on their child's performance at school is for most specifications significantly negative. It means that if parents are more interested in school outcomes, the probability of being retained decreases. For some specifications, the dummy variable for the highest education category of the mother is significant and negative.

In order to evaluate the common support assumption the density of estimated propensity scores by treatment status are drawn for all groups (see Figures from B 1 to B 12). The propensity score graphs do not exhibit a significant common support problem. Nevertheless, we estimate the ATEs twice for each sample. First, we do not apply any common support correction and second we use minima-maxima comparison (see Frölich (2004), Imbens (2004), Imbens and Wooldridge (2007), and Caliendo and Kopeinig (2008)). Minima-maxima comparison is simply discarding the control observations with propensity scores below the minimum propensity score of the treated group and discarding treated observations with propensity scores above the maximum propensity score of the control group.

The estimation results are summarized in Table 2 and 3. Table 2 shows the results for the sample without any restrictions and Table 3 shows the results for the sample of students without previous retention. we estimate the ATE of grade retention for the entire samples and for the subsamples by gender. The estimates of causal effect on high school completion are summarized in the upper panel and the estimates of

causal effect on academic grades in the lower panel of Table 2 and 3. The effects are estimated using Doubly Robust Method (DR) (Equation 2.12), weighting by propensity score (PS) (Equation 2.5) and regression (REG) (Equation 2.3) which are outlined in Section 2. For the regression and DR method, the mean functions of the outcome variables are chosen properly according to the features of the outcome variables. The mean function of the binary outcome variable ABI is specified as in Equation 2.9. For the outcome variables MAT and GER, Equation 2.10 is chosen as the mean function. The mean of the last outcome variable GPA is chosen as in Equation 2.8. The control variables are the same as in the propensity score specifications. For each sample, there are two different sets of estimates; column (a) and (b). Column (a) shows the estimation results without applying any common support correction. For the estimates in column (b), we apply minima-maxima comparison to determine the common support. The standard errors are calculated using the asymptotic variance formulas and reported in parentheses.

From Table 2, we see that the effect of grade retention on the probability of completion of upper secondary school for the overall sample is negative according to the DR and REG estimates. The negative effect is higher in magnitude for females than for the entire sample, whereas the effect seems to be positive for males. For all three samples, PS estimates are insignificant. Applying common support restriction only slightly affects the estimates. For the other three outcomes, the estimates by each method are significantly positive for each sample with two exceptions. The PS estimates of the ATE on MAT for females is insignificant with and without common support restriction. The PS estimates of the ATE on GPA for females are insignificant without common support restriction. In the German educational system, grades between 1 and 6 are assigned, where 1 is the best grade and 6 is the worst grade. Therefore, positive estimates of ATE imply a worsening effect on grades. We see that the estimates based on different methods are most of the time very close to each other. The estimates based on DR and REG methods are almost for each case highly significant whereas the PS estimates are sometimes insignificant. It is known that the variance of PS estimates are affected largely by very high and low propensity scores (see for example Khan and Tamer (2007)).

Table 3 shows the estimation results for the students who only experienced grade retention after 10^{th} grade. The results are very similar to the previous Table, except that the effect of grade retention on the probability of graduating from high

school for male students is significantly negative. Moreover, the estimates are larger in magnitude compared to the previous results. As in previous results, regardless of which method is used the estimates are very close for the same outcome variable. This result should give us some confidence about our model specifications. The negative effect of grade retention on high school completion is higher for boys than girls. Furthermore, the treatment effects on different school grades are also higher for boys than girls. It seems like boys are more negatively affected by grade retention than girls. All in all, our empirical results suggest that grade retention as a school intervention tool does not provide any improvement on average, but has rather worsening effects for students.

Table 2: Estimated ATE's for the main sample without restrictions according to previous retention.

Outcome	Method	Full Sample		Female		Male	
		(a)	(b)	(a)	(b)	(a)	(b)
ABI	**DR**	-0.010^{**}	-0.012^{**}	-0.043^{***}	-0.048^{***}	0.017^{***}	0.015^{**}
		(0.006)	(0.005)	(0.013)	(0.012)	(0.007)	(0.007)
	PS	0.002	0.001	-0.026	-0.033	0.018	0.018
		(0.024)	(0.024)	(0.040)	(0.039)	(0.029)	(0.028)
	REG	-0.006	-0.007^{**}	-0.033^{***}	-0.044^{***}	0.014^{***}	0.011^{**}
		(0.004)	(0.004)	(0.006)	(0.007)	(0.005)	(0.005)
number of observations		2726	2711	1257	1200	1469	1436
number of treated		520	519	201	196	319	316
number of untreated		2206	2192	1056	1004	1150	1120
MAT	**DR**	0.266^{***}	0.262^{***}	0.118^{***}	0.123^{***}	0.395^{***}	0.377^{***}
		(0.016)	(0.016)	(0.046)	(0.045)	(0.024)	(0.024)
	PS	0.255^{***}	0.257^{***}	-0.025	0.009	0.405^{***}	0.416^{***}
		(0.067)	(0.066)	(0.105)	(0.104)	(0.083)	(0.083)
	REG	0.271^{***}	0.273^{***}	0.104^{***}	0.109^{***}	0.401^{***}	0.383^{***}
		(0.010)	(0.010)	(0.022)	(0.023)	(0.016)	(0.016)
GER	**DR**	0.296^{***}	0.298^{***}	0.356^{***}	0.358^{***}	0.314^{***}	0.286^{***}
		(0.014)	(0.013)	(0.050)	(0.047)	(0.018)	(0.017)
	PS	0.295^{***}	0.301^{***}	0.225^{**}	0.264^{***}	0.326^{***}	0.341^{***}
		(0.058)	(0.057)	(0.102)	(0.100)	(0.072)	(0.069)
	REG	0.301^{***}	0.300^{***}	0.363^{***}	0.363^{***}	0.308^{***}	0.284^{***}
		(0.008)	(0.008)	(0.022)	(0.022)	(0.013)	(0.013)
GPA	**DR**	0.220^{***}	0.219^{***}	0.177^{***}	0.182^{***}	0.256^{***}	0.242^{***}
		(0.009)	(0.009)	(0.028)	(0.027)	(0.010)	(0.010)
	PS	0.213^{***}	0.219^{***}	0.100	0.135^{*}	0.256^{***}	0.274^{***}
		(0.041)	(0.039)	(0.077)	(0.075)	(0.047)	(0.044)
	REG	0.225^{***}	0.224^{***}	0.189^{***}	0.192^{***}	0.258^{***}	0.245^{***}
		(0.004)	(0.005)	(0.013)	(0.013)	(0.006)	(0.006)
number of observations		1643	1620	686	672	957	922
number of treated		303	299	105	105	198	197
number of untreated		1340	1321	581	567	759	725

The standard errors are calculated as explained in Section 2 and reported in parentheses under the estimates. Column (a) and (b) report the estimates without and with common support restriction respectively. *, **, ***: significant at 10 %, 5 %, 1%

Table 3: Estimated ATE's for the samples without previous retention

Outcome	Method	Full Sample		Female		Male	
		(a)	(b)	(a)	(b)	(a)	(b)
ABI	**DR**	-0.072^{***}	-0.073^{***}	-0.065^{***}	-0.071^{***}	-0.088^{***}	-0.089^{***}
		(0.006)	(0.006)	(0.013)	(0.013)	(0.008)	(0.009)
	PS	-0.062^{**}	-0.059^{**}	-0.049	-0.055	-0.078^{**}	-0.075^{**}
		(0.028)	(0.028)	(0.042)	(0.041)	(0.034)	(0.033)
	REG	-0.075^{***}	-0.075^{***}	-0.064^{***}	-0.068^{***}	-0.090^{***}	-0.092^{***}
		(0.004)	(0.004)	(0.009)	(0.009)	(0.006)	(0.006)
number of observations		1748	1738	866	842	882	850
number of treated		377	377	160	158	217	212
number of untreated		1371	1361	706	684	665	638
MAT	**DR**	0.351^{***}	0.348^{***}	0.143^{***}	0.148^{***}	0.549^{***}	0.542^{***}
		(0.023)	(0.023)	(0.050)	(0.050)	(0.032)	(0.031)
	PS	0.368^{***}	0.363^{***}	0.085	0.108	0.502^{***}	0.541^{***}
		(0.079)	(0.078)	(0.114)	(0.113)	(0.103)	(0.103)
	REG	0.365^{***}	0.367^{***}	0.140^{***}	0.143^{***}	0.546^{***}	0.546^{***}
		(0.012)	(0.012)	(0.027)	(0.028)	(0.018)	(0.019)
GER	**DR**	0.344^{***}	0.342^{***}	0.330^{***}	0.331^{***}	0.402^{***}	0.359^{***}
		(0.024)	(0.024)	(0.052)	(0.051)	(0.026)	(0.024)
	PS	0.385^{***}	0.382^{***}	0.323^{***}	0.348^{***}	0.367^{***}	0.393^{***}
		(0.069)	(0.069)	(0.118)	(0.116)	(0.084)	(0.078)
	REG	0.365^{***}	0.365^{***}	0.352^{***}	0.351^{***}	0.415^{***}	0.371^{***}
		(0.011)	(0.011)	(0.027)	(0.027)	(0.021)	(0.020)
GPA	**DR**	0.243^{***}	0.241^{***}	0.169^{***}	0.170^{***}	0.288^{***}	0.273^{***}
		(0.014)	(0.013)	(0.027)	(0.026)	(0.013)	(0.013)
	PS	0.270^{***}	0.267^{***}	0.199^{**}	0.219^{***}	0.248^{***}	0.284^{***}
		(0.047)	(0.046)	(0.083)	(0.081)	(0.052)	(0.048)
	REG	0.255^{***}	0.256^{***}	0.197^{***}	0.195^{***}	0.292^{***}	0.278^{***}
		(0.006)	(0.006)	(0.016)	(0.016)	(0.009)	(0.009)
number of observations		1248	1242	546	536	702	662
number of treated		227	225	84	84	143	141
number of untreated		1021	1017	462	452	559	521

The standard errors are calculated as explained in Section 2 and reported in parentheses under the estimates. Column (a) and (b) report the estimates without and with common support restriction respectively. *, **, ***: significant at 10 %, 5 %, 1%

3.2 Grade Retention as an Instrumental Variable

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 10^{th} 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). The tests and interviews were conducted at three different points in time: during the 10^{th} grade (1970), at the age of 30 (1984) and at the age of 43 (1997). The 10^{th} 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 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 10^{th} grade, whereas the binary policy instrument is grade retention at 10^{th} 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.[12] To claim that the identifying assumptions in Section 2.2 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. Given that many recent papers like Heckman *et al.* (2006), Carneiro *et al.* (2007), Heineck and Anger (2008), 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

[12]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.

10^{th} 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 8.

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 9 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 10^{th} 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 9 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 10^{th} 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 9, 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 2.16, 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 10 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 2.6) by comparing the distributions (histograms) of the estimated instrument propensity scores by instrumental variable as suggested in Lechner (2010). Figure C 13 shows that there is no common support problem.[13]

In order to apply the doubly robust or the regression method, we also specify the

[13]The observations shown as off support in Figure C 13 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.

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 4 presents the LATE estimates and the standard errors based on the different methods.[14]

Table 4: 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 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.

[14]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.

4 Conclusion

In this paper, we investigate the causal effect of grade retention on different school outcomes, such as completion of upper secondary school, final grades in math and German as well as the average final grade. The effect of grade retention is an important research topic since at least four decades. The results from previous research are somehow controversial. The literature provides evidence for both negative and positive effects. The methods used for the analysis of the effects range from simple group comparisons to sophisticated econometric modeling. Here, we estimate the effect using a potential outcome framework applying econometric evaluation methods inverse propensity score weighting, regression adjustment and a combination of these two methods. Inverse propensity score weighting estimates are inconsistent if the propensity score is wrongly specified and regression adjustment estimates are inconsistent if the mean function is wrongly specified. Hence, a combination of these two methods gives the researcher some protection against misspecification. The resulting estimator of the ATE is consistent even if only one of the models is correctly specified. An important drawback is that the main underlying assumption, CIA, which provides the identification of the average treatment effect is not testable. As most researchers who uses identification under CIA, we also argue that the rich set of control variables we use should be enough to satisfy the CIA assumption approximately.

The propensity score estimation results are consistent with much of the existing empirical research on determinants of grade retention. The estimates of the ATE on different outcomes are very close to each other regardless of which of the three methods is chosen. The estimates show that grade retention has a worsening effect on the students' educational achievement. It increases drop-out rate from upper secondary school significantly, and decreases the individual grades in math and German as well as the average final grade. The worsening effect is larger for boys than for girls. Given that grade retention is thought as an intervention tool to improve the educational achievement, our result do not support that this intervention achieves that goal. This result coincides with other empirical results from the US and Canada (see for example Jimerson (1999) and Guevremont, Roos, and Brownell (2007)) and implies the necessity of different approaches to improve the educational achievement.

In the second part, we propose a combination of weighting and regression methods of the LATE estimation which is doubly robust. The weighting method alone re-

quires 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 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. (2004): "A Note on the Role of the Propensity Score for Estimating Average Treatment Effects," *Econometric Reviews*, 23(2), 167 – 174.

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

Guevremont, A., N. P. Roos, and M. Brownell (2007): "Predictors and Consequences of Grade Retention: Examining Data from Manitoba, Canada," *Canadian Journal of School Psychology*, 22, 50–67.

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.

Heineck, G., and S. Anger (2008): "The Returns to Cognitive Abilities and Personality Traits in Germany," Discussion paper, German Institute for Economic Research.

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.

Horvitz, D. G., and D. J. Thompson (1952): "A Generalization of Sampling Without Replacement From a Finite Universe," *Journal of the American Statistical Association*, 47, 663–685.

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.

Imbens, G., and J. Wooldridge (2007): "What is New in Econometrics, Lecture 1, Estimation of Average Treatment Effects under Unconfoundedness," NBER Lectures.

Jackson, G. B. (1975): "The Research Evidence on the Effects of Grade Retention," *Review of Educational Research*, 45, 613–635.

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.

Jimerson, S. R. (1999): "On the Failure of Failure: Examining the Association Between Early Grade Retention and Education and Employment Outcomes During Late Adolescence," *Journal of School Psychology*, 37, 243272.

——— (2001): "Meta-Analysis of Grade Retention Research: Implications for Practice in the 21st Century," *School Psychology Review*, 30, 420438.

Khan, S., and E. Tamer (2007): "Irregular Identification, Support Conditions, and Inverse Weight Estimation," Discussion paper, Unpublished manuscript, Northwestern University 2009.

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

McCoy, A. R., and A. J. Reynolds (1999): "Grade Retention and School Performance: An Extended Investigation," *Journal of School Psychology*, 37, 273–298.

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.

Rosenbaum, P. R. (1987): "Model-Based Direct Adjustment," *Journal of the American Statistical Association*, 82, 387–394.

Rosenbaum, P. R., and D. B. Rubin (1983): "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70, 41 – 55.

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.

Scharfstein, D. O., A. Rotnitzky, and J. M. Robins (1999): "Adjusting for nonignorable drop-out using semiparametric nonresponse models (with discussion)," *Journal of the American Statistical Association*, 94, 1096–1120.

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.

——— (2009): "Average Treatment Effect Estimation: Unconfounded Treatment Assignment," Unpublished Manuscript, Michigan State University.

A Appendix

A.1 Proofs

Proof 1 [15]

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

The first three equations follow from the definition of the potential outcome and the potential treatment status. The fourth equation follows from Assumption A 2.3.

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

Therefore,

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

Due to Assumption A 2.5, the second term in the first equation is equal to zero. Moreover, since $\mathrm{E}\left[D|\, X, Z=0\right] = \Pr\left[D=1\,|X, Z=0\right] = P[\textit{always takers}|X] + P[\textit{defiers}|X]$ *and* $\mathrm{E}\left[D|\, X, Z=1\right] = \Pr\left[D=1\,|X, Z=1\right] = P[\textit{always takers}|X] + P[\textit{compliers}|X]$*, the relative size of the subpopulation of compliers is identified as:*

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

Therefore,

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

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

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

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

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:*

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

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

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

Thus,

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

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. (2.26):

$$\hat{V} = \frac{1}{N}\sum_i \psi(W_i,\hat{\theta}^w)\psi(W_i,\hat{\theta}^w)'$$

$$\hat{A} = \frac{1}{N}\sum_i \frac{\partial\psi(W_i,\hat{\theta}^w)}{\partial\theta'}$$

$$= \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}$$

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

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_i X_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_i X_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_i X_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_i X_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_i X_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_i X_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_i X_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_i X_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_i X_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: Definition of the Variables used for estimation of the ATE

Variable	Definition
ABI	Dummy, 1 if upper secondary school degree held (Abitur)
MAT	Grade in math in the last year of upper secondary school between 1-6, 1 is the best grade
GER	Grade in German in the last year of upper secondary school between 1-6, 1 is the best grade
RET	Dummy, 1 if a grade is repeated at least once in the school year 1970/71 or later
SHNR	School number
FEMALE	Dummy, 1 if female
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.
EDU_MOT	Categorical variable for educational attainment of the mother from 1-4
EDU_MOTj	Dummy, 1 if EDU_MOT=j for $j = 1, 2, 3, 4$
EDU_FAT	Categorical variable for educational attainment of the father from 1-4
EDU_FATj	Dummy, 1 if EDU_FAT=j for $j = 1, 2, 3, 4$
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
EMP_MOT1	Dummy, 1 if the mother is employed during the survey (EMP_MOT=1)
EMP_MOT2	Dummy, 1 if the mother is unemployed, but was employed before the survey (EMP_MOT=2)
EMP_MOT3	Dummy, 1 if the mother is out of labour force (EMP_MOT=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
AGEMOTj	Dummy, 1 if AGEMOT=j
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
WISHj	Dummy, =1 if WISH=j
PR_RET	Dummy, 1 if a grade is repeated at least once before the school year 1969/70
DILIG	Measure of attributing success to diligence 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)

Source: Dataset Gymnasiastenstudie, own definitions

Table B2: Summary Statistics of Unrestricted Sample (Sample 1)

Variable	Mean	Std Dev	Minimum	Maximum
ABI	0.64	0.48	0	1
RET	0.19	0.39	0	1
FEMALE	0.46	0.50	0	1
AGE	15.41	0.90	13	19
IQ	40.72	8.94	12	70
EDU_MOT	4.19	3.50	1	13
EDU_VAT	5.86	4.26	1	13
HHINC	4.50	2.06	1	9
EMP_MOT	2.02	0.70	1	3
PARINT1	0.64	0.48	0	1
PARINT2	0.61	0.49	0	1
PARINT3	0.75	0.43	0	1
INTSCHOOL	0.67	0.30	0	1
AGE_MOT	3.61	1.18	1	9
PR_RET	0.36	0.48	0	1
WISH	2.62	1.56	1	5
DILIG	4.12	1.04	0	5
ABIL	3.51	1.08	0	5
Number of Observations			2726	

Sample: Sample without restrictions on previous grade retention

Table B3: Summary Statistics of Sample 2

Variable	Mean	Std	Minimum	Maximum
GPA	2.97	0.54	1.08	4.10
MAT	3.48	1.08	1	6
GER	3.33	0.85	1	5
RET	0.18	0.39	0	1
FEMALE	0.42	0.49	0	1
AGE	15.19	0.82	13	19
IQ	41.78	9.12	15	70
EDU_MOT	4.36	3.60	1	13
EDU_VAT	6.10	4.31	1	13
HHINC	4.57	2.06	1	9
EMP_MOT	2.03	0.69	1	3
PARINT1	0.64	0.48	0	1
PARINT2	0.64	0.48	0	1
PARINT3	0.78	0.41	0	1
INTSCHOOL	0.69	0.30	0	1
AGE_MOT	3.62	1.18	1	9
PR_RET	0.24	0.43	0	1
WISH	2.15	1.32	1	5
DILIG	4.09	1.05	0	5
ABIL	3.51	1.09	0	5
Number of Observations			1643	

Sample: Graduates from upper secondary school. Sample 1 restricted by ABI=1

Table B4: Summary Statistics of Sample 3

Variable	Mean	Std. Dev.	Minimum	Maximum
ABI	0.75	0.43	0	1
RET	0.22	0.41	0	1
FEMALE	0.50	0.50	0	1
AGE	15.03	0.71	13	19
IQ	40.85	9.08	13	70
EDU_MOT	4.15	3.55	1	13
EDU_VAT	5.76	4.27	1	13
HHINC	4.41	2.07	1	9
EMP_MOT	2.03	0.69	1	3
PARINT1	0.64	0.48	0	1
PARINT2	0.63	0.48	0	1
PARINT3	0.77	0.42	0	1
INTSCHOOL	0.68	0.30	0	1
AGE_MOT	3.53	1.16	1	9
WISH	2.45	1.50	1	5
DILIG	4.11	1.03	0	5
ABIL	3.57	1.05	0	5
Number of Observations			1748	

Sample: Students without previous grade retention. Sample 1 restricted by PR_RET=0

Table B5: Summary Statistics of Sample 4

Variable	Mean	Std. Dev.	Minimum	Maximum
GPA	2.92	0.54	1.08	4.00
MAT	3.36	1.09	1	6
GER	3.26	0.87	1	5
RET	0.18	0.39	0	1
FEMALE	0.44	0.50	0	1
AGE	14.96	0.69	13	19
IQ	41.79	9.22	15	70
EDU_MOT	4.26	3.62	1	13
EDU_VAT	5.92	4.29	1	13
HHINC	4.46	2.06	1	9
EMP_MOT	2.03	0.68	1	3
PARINT1	0.65	0.48	0	1
PARINT2	0.79	0.41	0	1
PARINT3	2.12	1.30	1	5
INTSCHOOL	0.69	0.30	0	1
AGE_MOT	3.56	1.17	1	9
WISH	2.12	1.30	1	5
DILIG	4.10	1.04	0	5
ABIL	3.56	1.06	0	5
Number of Observations			1248	

Sample: Graduates from upper secondary school without previous grade retention. Sample 3 restricted by ABI=1

Table B6: Propensity Score Estimation Results for Sample 1 and Sample 2

	DATA 1			DATA 2		
	(a)	(b)	(c)	(d)	(e)	(f)
Variable	Full sample	Female	Male	Full sample	Female	Male
Constant	3.060***	3.067	2.506*	3.750**	2.870	3.575*
	1.242	2.091	1.555	1.680	3.101	2.073
SHNR	−0.006**	−0.009**	-0.004	−0.009***	−0.015***	-0.006
	0.003	0.004	0.003	0.003	0.006	0.004
FEMALE	−0.534***	–	–	−0.604***	–	–
	0.110			0.146		
AGE	-0.093	-0.103	-0.070	0.013	0.079	0.013
	0.070	0.122	0.087	0.093	0.179	0.111
IQ	−0.034***	−0.039***	−0.031***	−0.039***	−0.063***	−0.027***
	0.006	0.010	0.008	0.008	0.014	0.010
EDU_MOT2	-0.060	-0.039	-0.046	-0.127	-0.020	-0.206
	0.141	0.218	0.188	0.188	0.312	0.241
EDU_MOT3	-0.105	-0.136	-0.140	-0.272	-0.455	-0.309
	0.207	0.325	0.274	0.278	0.482	0.354
EDU_MOT4	-0.308	0.262	−0.777**	-0.453	0.370	-1.045
	0.243	0.358	0.337	0.319	0.474	0.445
EDU_VAT2	0.257*	0.385*	0.122	0.161	0.091	0.158
	0.155	0.230	0.216	0.214	0.343	0.279
EDU_VAT3	0.102	0.202	0.052	0.203	0.308	0.170
	0.164	0.258	0.216	0.210	0.361	0.264
EDU_VAT4	0.217	0.074	0.348	0.324	0.261	0.472
	0.190	0.302	0.249	0.249	0.403	0.322
HHINC	-0.036	−0.088*	0.000	-0.064	−0.132*	-0.028
	0.031	0.050	0.041	0.042	0.072	0.053
EMP_MOT1	0.315**	0.254	0.326*	0.305	0.141	0.318
	0.144	0.244	0.183	0.190	0.344	0.235
EMP_MOT2	0.069	0.158	0.035	-0.036	-0.020	-0.017
	0.125	0.204	0.161	0.161	0.274	0.204
INTSCHOOL	−0.384**	−0.852***	-0.070	-0.247	-0.586	-0.066
	0.169	0.277	0.219	0.219	0.379	0.275
AGEMOT2	-0.451	-0.557	-0.386	-0.590	-0.325	-0.632
	0.420	0.714	0.527	0.538	1.147	0.638
AGEMOT3	−0.671*	-0.538	-0.805	-0.570	-0.115	-0.734
	0.410	0.700	0.514	0.519	1.118	0.614
AGEMOT4	−0.735*	-0.696	-0.825	-0.733	-0.155	-0.993
	0.412	0.706	0.516	0.522	1.122	0.619
AGEMOT5	−0.741*	-0.650	-0.833	-0.874	-0.433	-1.019
	0.432	0.733	0.543	0.550	1.170	0.650
AGEMOT6	−1.225***	-0.840	−1.509**	−1.132*	-0.401	−1.385*
	0.492	0.805	0.635	0.612	1.232	0.741
AGEMOT7	−1.458*	-0.801	-1.794	-1.540	(omitted)	-1.483
	0.865	1.284	1.188	1.228		1.272
AGEMOT8	−1.612*	(omitted)	-1.424	-2.035	(omitted)	-1.775
	0.871		0.959	1.185		1.256
PR_RET	−0.414***	−0.439*	−0.414***	-0.037	-0.200	-0.002
	0.133	0.230	0.167	0.179	0.329	0.219
WISH1	0.422**	0.684***	0.176	−1.509***	−1.276**	−1.876***
	0.190	0.267	0.277	0.391	0.538	0.636
WISH2	0.537***	0.932***	0.191	−1.308***	-0.844	−1.768***
	0.204	0.285	0.301	0.402	0.553	0.653
WISH3	0.644***	0.482	0.664*	−1.272***	−1.368*	−1.336*
	0.256	0.426	0.347	0.465	0.737	0.706
WISH4	0.788***	0.829***	0.691***	−1.105***	−1.045**	−1.383**
	0.187	0.265	0.275	0.390	0.538	0.637
DILIG	−0.076*	-0.086	-0.071	-0.129	-0.098	−0.145**
	0.047	0.081	0.059	0.061	0.115	0.073
ABIL	−0.104**	-0.059	−0.138**	−0.111*	-0.085	−0.153**
	0.046	0.077	0.058	0.059	0.103	0.074
No. of Obs.	2726	1249	1469	1643	680	957
Log-likelihood	-1258.31	-515.10	-729.95	-740.46	-267.99	-461.91
LR chi2(k)	140.22	71.92	77.50	89.89	49.20	51.97

The standard errors are reported in parentheses under the estimates. *, **, ***: significant at 10 %, 5 %, 1%

Table B7: Propensity Score Estimation Results for Different Samples data2

	DATA 3			DATA 4		
	(a)	(b)	(c)	(d)	(e)	(f)
Variable	Full sample	Female	Male	Full sample	Female	Male
Constant	1.049**	-1.096	2.550	2.290	-1.913	4.209
	1.545	2.383	2.117	2.061	3.578	2.730
SHNR	−0.006**	-0.006	-0.006	−0.009**	-0.012	-0.008
	0.003	0.005	0.004	0.004	0.006	0.005
FEMALE	−0.561***	–	–	−0.624***	–	–
	0.131			0.168		
AGE	0.062	0.154	-0.004	0.111	0.362*	0.025
	0.087	0.140	0.116	0.113	0.208	0.139
IQ	−0.034***	−0.030***	−0.037***	−0.035***	−0.055***	−0.027**
	0.007	0.011	0.010	0.009	0.015	0.011
EDU_MOT2	-0.026	0.010	-0.051	-0.118	-0.028	-0.186
	0.172	0.251	0.243	0.223	0.355	0.295
EDU_MOT3	-0.061	-0.099	-0.071	-0.255	-0.488	-0.156
	0.256	0.368	0.370	0.329	0.529	0.446
EDU_MOT4	-0.435	-0.033	−0.817*	-0.634	0.123	−1.151**
	0.302	0.431	0.431	0.391	0.566	0.568
EDU_VAT2	0.395**	0.492*	0.271	0.360	0.263	0.375
	0.185	0.260	0.274	0.246	0.388	0.327
EDU_VAT3	0.088	0.230	0.040	0.359	0.465	0.352
	0.200	0.297	0.278	0.246	0.409	0.317
EDU_VAT4	0.272	0.171	0.356	0.444	0.518	0.443
	0.238	0.351	0.337	0.305	0.458	0.424
HHINC	-0.013	-0.069	0.040	-0.017	-0.077	0.024
	0.037	0.057	0.051	0.048	0.083	0.062
EMP_MOT1	0.544***	0.487*	0.570	0.479**	0.450	0.466*
	0.175	0.280	0.232	0.221	0.396	0.276
EMP_MOT2	0.173	0.259	0.157	0.040	0.241	-0.038
	0.153	0.236	0.205	0.191	0.320	0.244
INTSCHOOL	−0.525***	−0.904***	-0.277	-0.391	−0.761*	-0.183
	0.201	0.314	0.273	0.253	0.427	0.326
AGEMOT2	-0.291	-0.834	0.160	-0.293	-0.517	-0.271
	0.534	0.739	0.786	0.715	1.163	0.934
AGEMOT3	-0.525	-0.791	-0.346	-0.317	-0.458	-0.373
	0.526	0.724	0.775	0.701	1.137	0.918
AGEMOT4	-0.658	-1.096	-0.368	-0.476	-0.600	-0.559
	0.529	0.735	0.778	0.704	1.146	0.922
AGEMOT5	-0.555	-1.001	-0.215	-0.382	-0.644	-0.370
	0.552	0.769	0.808	0.733	1.201	0.952
AGEMOT6	-0.661	-0.914	-0.469	-0.441	-0.637	-0.241
	0.605	0.831	0.892	0.784	1.258	1.032
AGEMOT7	-1.702	(omitted)	-0.989	-1.194	(omitted)	-0.967
	1.208		1.374	1.344		1.462
AGEMOT8	-1.028	(omitted)	-0.688	-1.197	(omitted)	-1.070
	0.973		1.168	1.304		1.460
WISH1	-0.166	0.279	−0.874**	−2.618***	−2.125***	−3.517***
	0.232	0.301	0.399	0.527	0.670	1.120
WISH2	0.182	0.625**	-0.537	−2.271***	−1.790***	−3.117***
	0.242	0.317	0.415	0.535	0.683	1.128
WISH3	0.342	0.193	0.018	−2.206***	−2.161***	−2.760**
	0.313	0.516	0.474	0.598	0.879	1.171
WISH4	0.309	0.562*	-0.230	−2.065***	−1.760***	−2.866***
	0.226	0.293	0.396	0.523	0.665	1.118
DILIG	−0.104*	-0.084	-0.118	-0.073	-0.013	-0.096
	0.057	0.091	0.075	0.073	0.135	0.089
ABIL	−0.123**	0.004	−0.219***	-0.088	0.044	−0.188**
	0.056	0.088	0.076	0.071	0.117	0.091
No. of Obs.	1748	860	882	1248	543	702
Log-likelihood	-863.10	-393.98	-457.55	-551.42	-214.12	-329.11
LR chi2(k)	96.54	38.40	69.10	80.88	39.57	51.50

The standard errors are reported in parentheses under the estimates. *, **, ***: significant at 10 %, 5 %, 1%

Table B8: Description of the variables used for estimation of the LATE

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 10^{th} 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 10^{th} 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 B9: 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 B10: 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

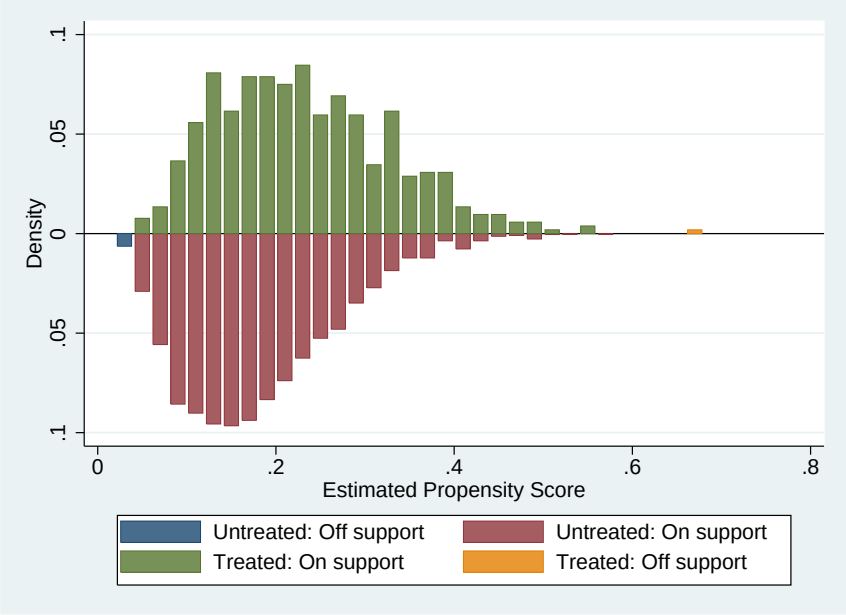

Figure C1: Density of Estimated Probability of Grade Retention for Sample 1. Estimation is based on specification given in Table A6, Col. (a)

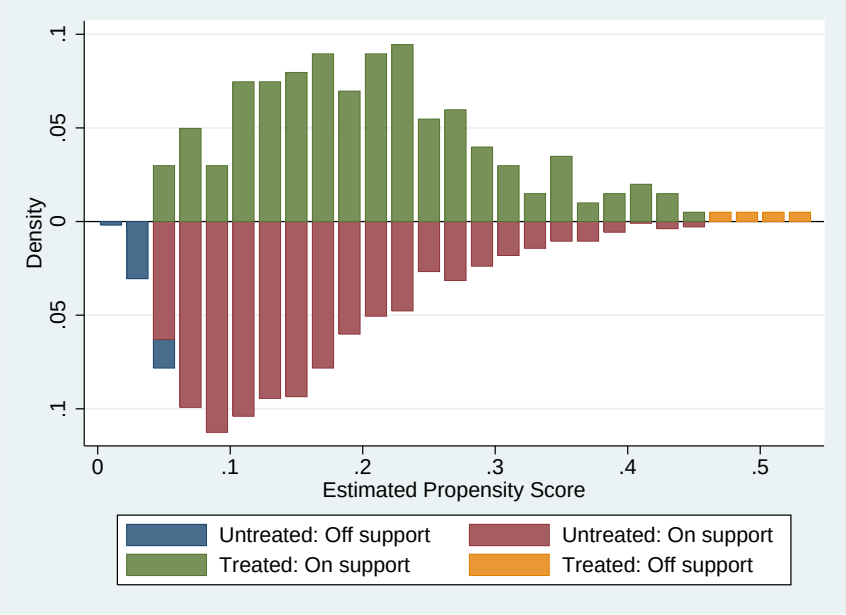

Figure C2: Density of Estimated Probability of Grade Retention for females of Sample 1. Estimation is based on specification given in Table A6, Col. (b)

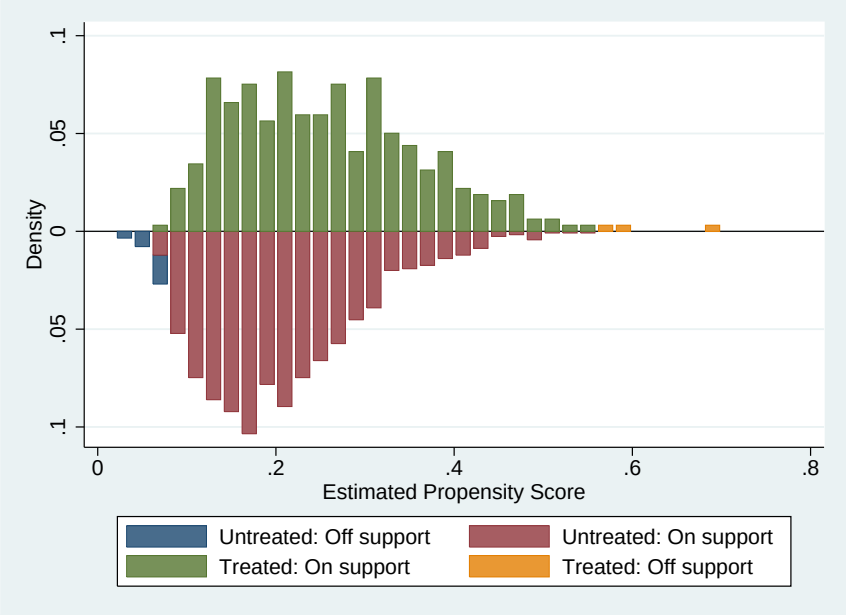

Figure C3: Density of Estimated Probability of Grade Retention for males of Sample 1. Estimation is based on specification given in Table A6, Col. (c)

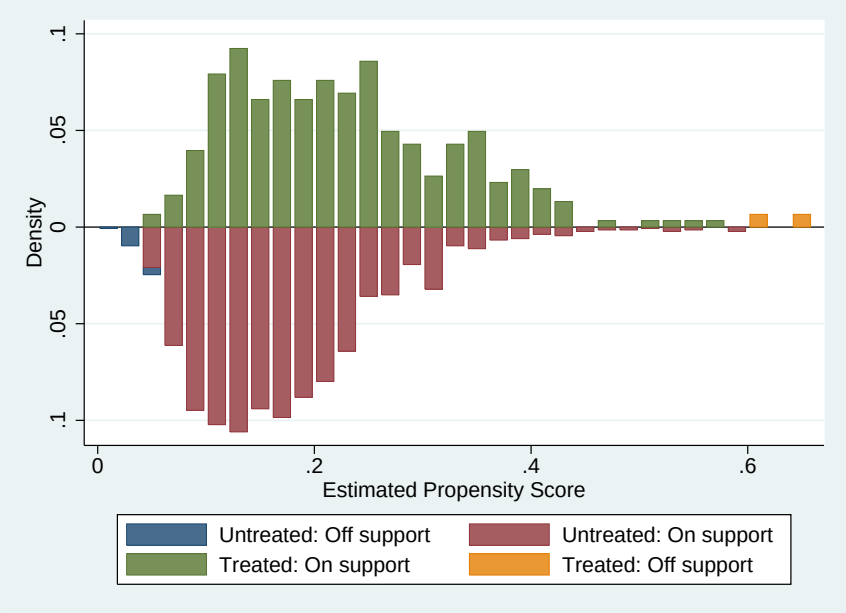

Figure C4: Density of Estimated Probability of Grade Retention for Sample 2. Estimation is based on specification given in Table A6, Col. (d)

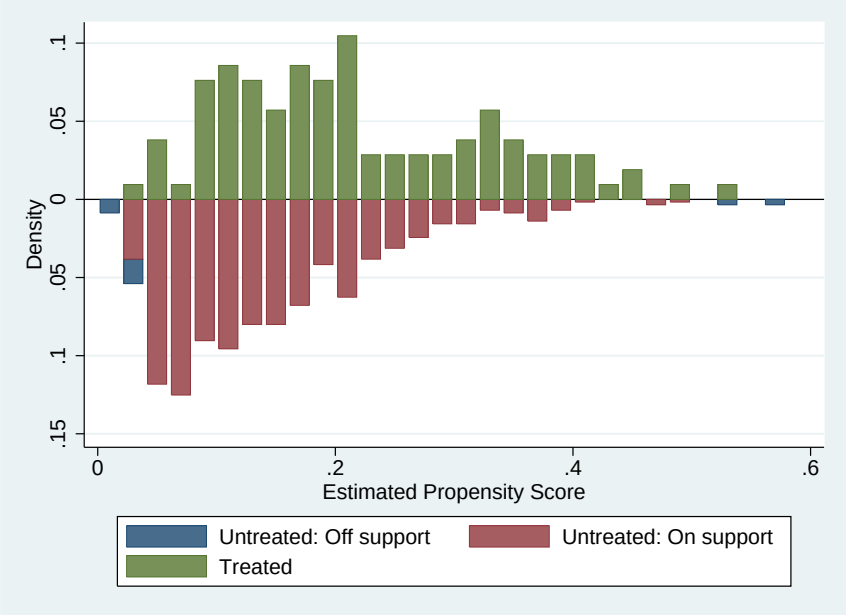

Figure C5: Density of Estimated Probability of Grade Retention for females of Sample 2. Estimation is based on specification given in Table A6, Col. (e)

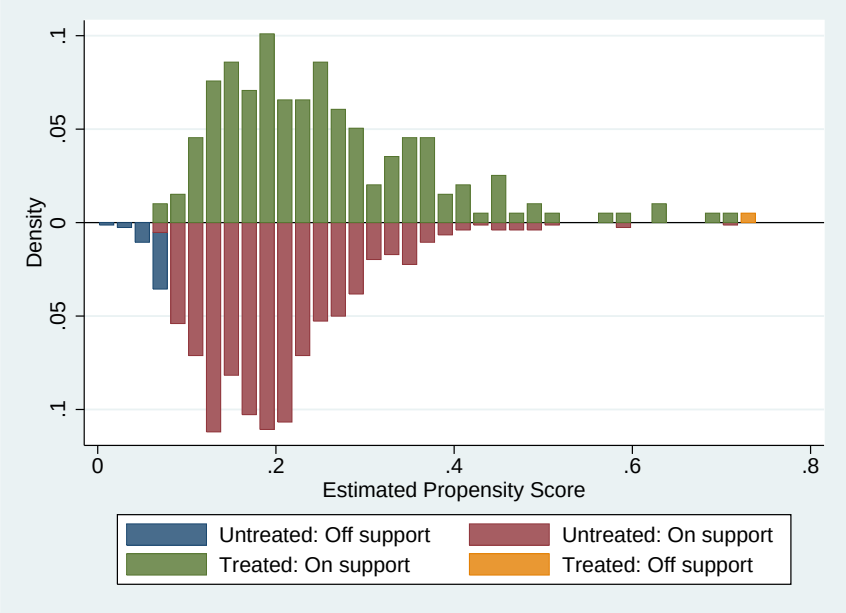

Figure C6: Density of Estimated Probability of Grade Retention for males of Sample 2. Estimation is based on specification given in Table A6, Col. (f)

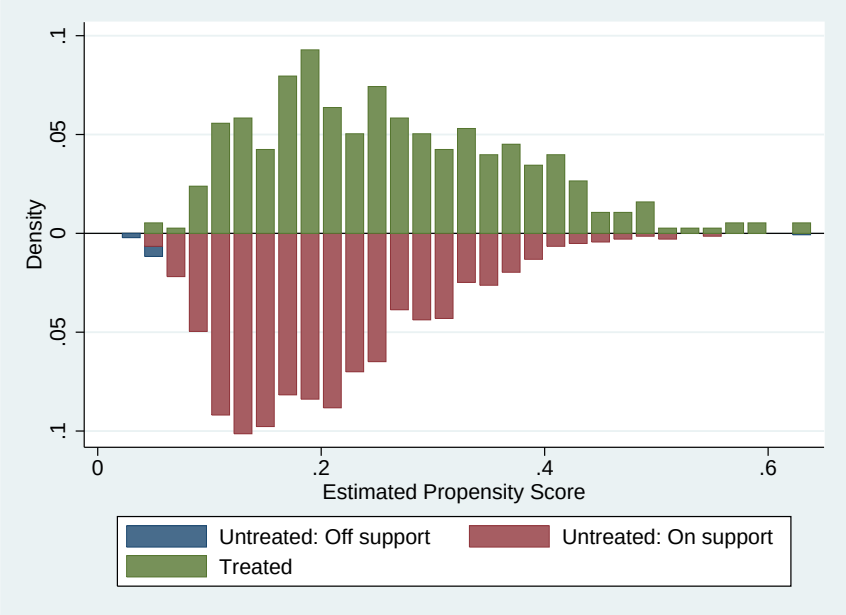

Figure C7: Density of Estimated Probability of Grade Retention for Sample 3. Estimation is based on specification given in Table A7, Col. (a)

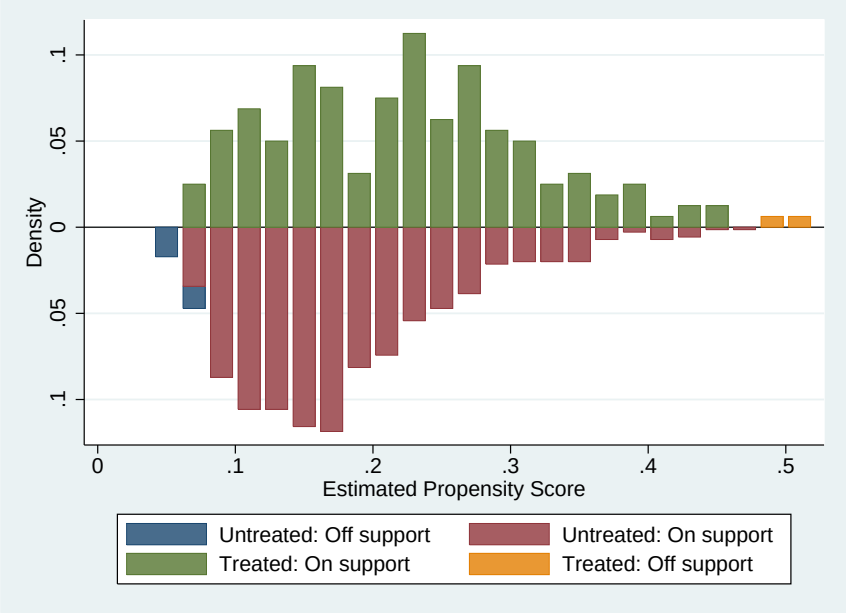

Figure C8: Density of Estimated Probability of Grade Retention for females of Sample 3. Estimation is based on specification given in Table A7, Col. (b)

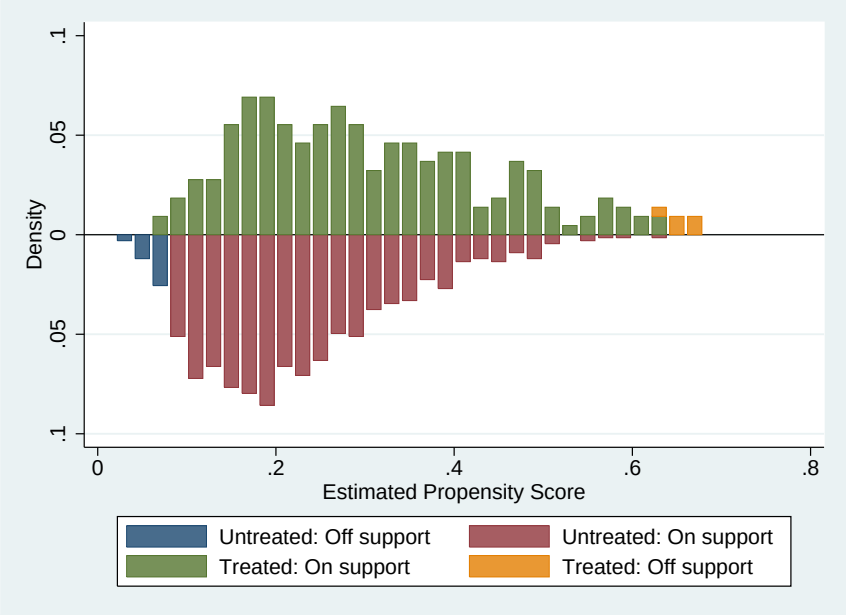

Figure C9: Density of Estimated Probability of Grade Retention for females of Sample 3. Estimation is based on specification given in Table A7, Col. (c)

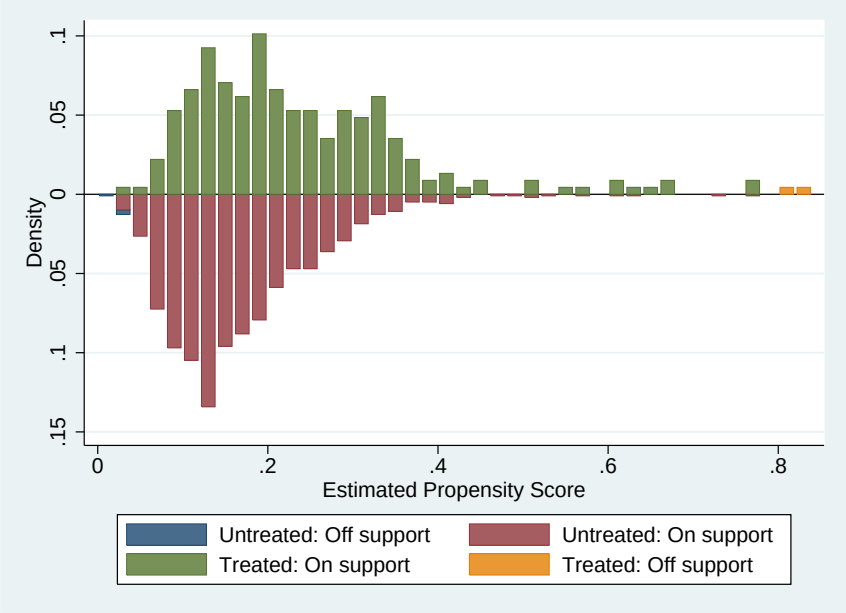

Figure C10: Density of Estimated Probability of Grade Retention for Sample 4. Estimation is based on specification given in Table A7, Col. (d)

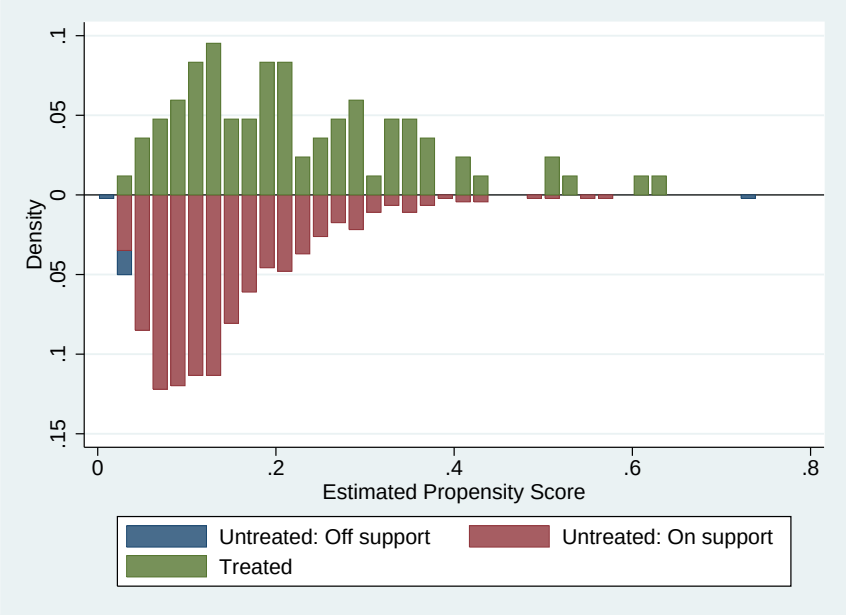

Figure C11: Density of Estimated Probability of Grade Retention for females of Sample 4. Estimation is based on specification given in Table A7, Col. (e)

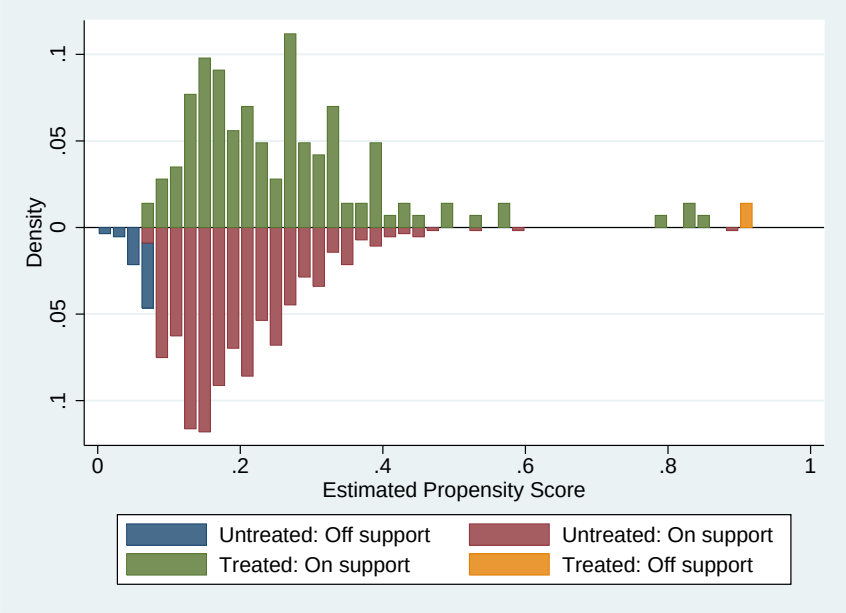

Figure C12: Density of Estimated Probability of Grade Retention for males of Sample 4. Estimation is based on specification given in Table A7, Col. (f)

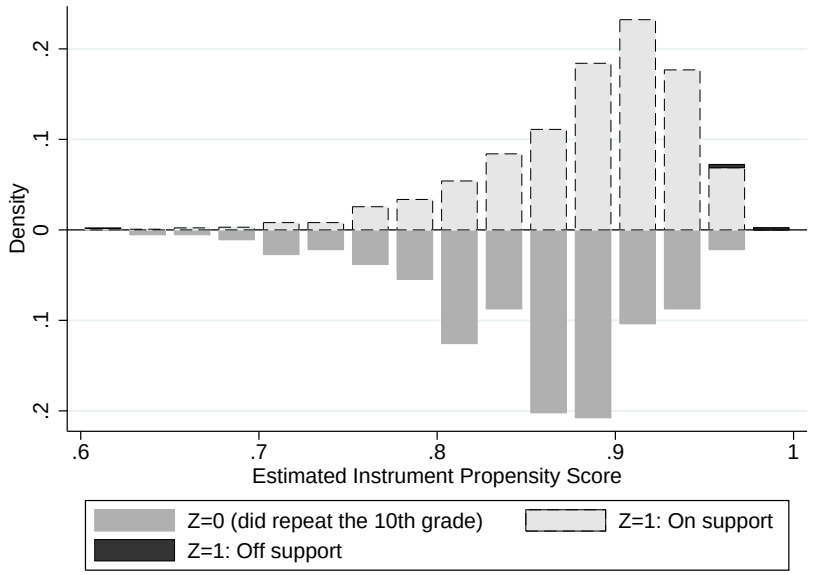

Figure C13: Density of Estimated Probability of Grade Retention. Estimation is based on specification given in Table B 10