The Impact of Retention on School Attainment: Local Average Treatment Effect(s) with a Multivalued Instrument

S. Derya Uysal*

Department of Economics

University of Munich

October 1, 2020

Abstract

We investigate the identification and estimation of different local average treatment effect (LATE) parameters that are defined in terms of a multivalued instrument. We extend the existing literature on conditionally valid multivalued instruments by explicitly defining two types of LATE and proposing (i) a weighted regression method and (ii) a normalized weighting estimator for both. Subsequently, we apply our proposed methods to estimate the causal effects on school attainment of retention at the end of 10th grade in Germany. We exploit the discontinuity induced by the rules for retention to construct our multivalued instrument. Our estimation indicates substantial treatment heterogeneity.

JEL classification: $C21, I21$

Keywords: Repeating a Grade; LATE; Multivalued Instrument; Treatment Effects; Double Robustness

*Department of Economics, University of Munich, Munich, Germany. Phone +49 89 2180 2224. email: derya.uysal@econ.lmu.de. The financial support by Deutsche Forschungsgemeinschaft through CRC TRR 190 is gratefully acknowledged.

1 Introduction

In this study, we examine the local average treatment effect (LATE) parameters with a multivalued instrument where the instrumental variable (IV) assumptions are valid only conditionally on covariates. We closely follow and combine the existing literature on LATE estimation (Abadie, 2003; Frölich, 2007; Tan, 2006; Uysal, 2011; Donald et al., 2014) and the estimation of treatment effect(s) in case of a multivalued treatment variable (Imbens, 2000; Lechner, 2001; Uysal, 2015). In particular, we explicitly study the (unconditional) LATEs, as well as the LATEs for subpopulations defined by the value of their instruments (conditional LATEs). Further, we discuss the estimation of these two types of LATE s through weighted regression and normalized weighting approaches. Our weighted regression approach possesses the double robustness property.[1] To the best of our knowledge, these specific extensions have not been explicitly discussed in the literature so far.

Subsequently, we apply the proposed methods to estimate the causal effect of being retained in 10th grade on the highest school degree achieved in Germany. For many countries, where being forced to repeat a grade is a common intervention, economists have recently provided insights on the impact of the retention on various outcomes. The empirical evidence on Germany is very scarce even though Germany is one of the countries with relatively high retention rates. The existing studies mostly provide descriptive analyses but do not uncover causal relations (see, for example, Ehmke et al., 2008; Demski & Liegmann, 2014). To fill this gap and provide further empirical evidence on the effectiveness of retaining students in a grade as an educational policy, we estimate the impact of being mandated to repeat the 10th grade on school leaving qualification, using data from Germany.

[1]The LATE estimator proposed by Belloni et al. (2017) is also doubly robust; however, its extension to a multivalued instrument has not been discussed.

One important reason for the lack of empirical evidence is undoubtedly the lack of data availability. For the analysis, we use the *Gymnasiastenstudie* data set, provided by the Central Archive for Empirical Social Research (CAESR, 2007). It provides us with detailed information on students' grades for all subjects taken at the end of the 10th grade and information on whether those grades led to promotion to the next grade level or retention. Additionally, due to the data's longitudinal feature, we observe the highest school degree obtained for the same individuals after ten years. The data also contain survey information on individual and family characteristics, as well as IQ test scores.

A common strategy to identify the causal effects of retention is the regression discontinuity.[2] The reason why this particular identification strategy is so prevalent in the retention literature is that retention is usually determined by a student's obtaining a score below a certain threshold. This institutional structure naturally leads to a discontinuity in the probability of retention. In many cases, a retention rule based on a test score gives rise to a sharp regression discontinuity design, that is, promotion is granted if the student's score is higher than a certain threshold (Jacob & Lefgren, 2004, 2009; Greene & Winters, 2007, 2009; Schwerdt et al., 2017; Eren et al., 2017). In some cases, the retention rule only causes a substantial change in the retention probability, as in Manacorda (2012). He exploits the discontinuity induced by a rule determining grade failure as more than three failed subjects. In Germany, a similar discontinuity is observed, with grade failure defined as failing more than one subject. The identification and estimation strategy used by Manacorda, however, does not seem to be suitable for our investigation. The first problem with

[2]The instrumental variable approach (Eide & Showalter, 2001; Élodie Alet et al., 2013; D'Haultfoeuille, 2010; Dong, 2010) and more structural approaches based on factor-analytic dynamic models (Fruehwirth et al., 2016; Gary-Bobo et al., 2016; Cockx et al., 2019) are also used in the retention literature.

our application is that our discrete running variable is very coarsely distributed over a very limited support (Lee & Lemieux, 2010).[3] Second, neither the classical regression discontinuity framework based on the continuity assumption nor the local randomization framework for regression discontinuity would be reasonable in our case since both rely on the notion that individuals on either side of the threshold are similar to each other except for their treatment status (see Cattaneo et al., 2020, chap. 1).However, in our application, we claim that individuals on either side of the threshold are comparable only after conditioning on their overall school performance. We formalize the identification and estimation strategy for our empirical question within the framework of LATE estimation with a multivalued instrumental variable.

This paper is structured as follows: In Section 1, we introduce the notation, define the treatment effects of interest, and review the identifying assumptions. Section 2 describes the estimation methods. Section 3 explains the data set and discusses the regulations for grade retention. In Section 4, the results are discussed in detail. Finally, Section 5 presents the conclusions.

2 Basic Framework and Identification

We adopt Rubin's (1974; 1977) potential outcome framework for causal inference and closely follow the notation of Tan (2006) and Frölich (2007). Consider N units drawn from a large population. For each individual i in the sample, where $i = 1, ..., N$, the quadruple (Y_i, D_i, Z_i, X_i) is observed. Y_i is the observed response variable, X_i denotes a vector of covariates, and D_i is the observed binary treatment status for

[3]Due to the same reason, the regression discontinuity approaches by Calonico et al. (2019) and Frölich & Huber (2019) are also not suitable for our investigation.

individual i:

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

Z_i is the instrument, which takes integer values between 1 and K.

The vast majority of studies dealing with an instrumental variable estimation within the potential outcome framework focus on the binary instrument case, with some notable exceptions. Tan (2006), for example, does not restrict the instrument to be binary in his discussion, and Frölich (2007) discusses the case of a non-binary instrument with bounded support. Here, we build upon their results and consider two types of treatment effects with a multivalued instrument. We also study the estimation of these effects in detail and offer extensions to existing methods.

The observed treatment variable is a function of the multivalued instrument Z_i and the potential treatment variables $D_i(z)$. For each individual there is a set of potential treatment variables $(D_i(1), \ldots, D_i(K))$. $D_i(z)$ denotes the treatment for each individual i, for which $Z_i = z$, where $z \in \mathcal{Z} = \{1, \ldots, K\}$. Formally, we can write the observed treatment variable as follows:

$$D_i = \sum_{k=1}^{K} \mathbb{1}\{Z_i = k\} D_i(k),$$

where $\mathbb{1}\{a\}$ is the indicator function, which returns one if a is true. Similarly, the potential outcome is denoted by $Y_i(d, z)$, where $d \in \{0, 1\}$ and $z \in \mathcal{Z}$; it is the value of the outcome that would occur if the treatment and instrument were equal to d and z, respectively. As in Imbens & Angrist (1994), one can partition the population with respect to the relation between the instrument and potential treatment into four

subpopulations: always takers, never takers, compliers, and defiers. Always takers, as the name suggests, always take the treatment, independent of the value of the instrument, that is, $D_i(z) = 1$ for all $z \in \mathcal{Z}$, and, similarly, never takers never take the treatment, $D_i(z) = 0$ for all $z \in \mathcal{Z}$. Compliers are those whose treatment status would change from zero to one if the value of the instrument Z_i were changed from z to z', that is, to a value that is more favorable for the treatment. Defiers are the ones who change their treatment status in the opposite direction. Defiers are usually assumed not to exist.[4] Under the assumption that defiers do not exist, and some additional assumptions discussed below, the effect of the treatment can be identified for compliers, the only subpopulation whose treatment can be manipulated by the instrument. In fact, in the case of a multivalued instrument, several LATEs can be identified for different pairs of (z, z'). The treatment effect for compliers is called the LATE and is formally given by the following expectation:

$$\tau_{LATE}(zz') = \mathrm{E}\left[Y_i(1) - Y_i(0) | \, D_i(z') > D_i(z)\right], \tag{2.1}$$

where $Y_i(1)$ and $Y_i(0)$ are the potential outcomes under the treatment ($D_i = 1$) and control ($D_i = 0$), respectively. This is the average treatment effect (ATE) for the subpopulation whose potential treatment status would switch to one if z is moved to z'. Note that this subpopulation might include individuals with any instrument value in $\mathcal{Z}$. If we are interested in the treatment effect for compliers whose instrument is equal to a specific instrument value, we can define another treatment effect parameter, which is the conditional version of the effect in (2.1), that is,

$$\tau_{LATE}(zz'|\mathcal{A}) = \mathrm{E}\left[Y_i(1) - Y_i(0) | \, D_i(z') > D_i(z), Z_i \in \mathcal{A}\right] \tag{2.2}$$

[4] de Chaisemartin (2017) discusses the identification and estimation of the LATE without ruling out the existence of defiers.

where $\mathcal{A}$ is a subset of $\mathcal{Z}$. For example, $\mathcal{A}$ could be $\{z, z'\}$, referring to the subpopulation of compliers with instrument levels equal to z or z'. In the case of a binary instrument, the two treatment effects $\tau_{LATE}(zz')$ and $\tau_{LATE}(zz'|zz')$ coincide. However, for a multivalued instrument, they do not only differ, but it might be the case that the conditional one is more policy-relevant. The argument is analogous to the one for the difference between the ATE and the average treatment effect on the treated (ATT).

We can further write the LATE for the treated (LATT) as follows:

$$\begin{aligned}
\tau_{LATT}(zz') &= \mathrm{E}\left[Y_i(1) - Y_i(0) | \, D_i(z') > D_i(z), D_i = 1\right] \\
&= \mathrm{E}\left[Y_i(1) - Y_i(0) | \, D_i(z') > D_i(z), Z_i \in \{l | l \in \mathcal{Z}, l \geq z'\}\right] \\
&= \tau_{LATE}(zz'|\mathcal{A}),
\end{aligned}$$

where $\mathcal{A} = \{l | l \in \mathcal{Z}, l \geq z'\}$. Since the potential treatment status would switch to one if z is moved to z' for compliers, treated compliers will have $Z_i \geq z'$. Therefore, the LATE for those with $Z_i \geq z'$ is equal to the LATT.

The earliest studies on the LATE obtained identification results under the assumption that the instrument is independent of the potential outcome and treatment variables (see, for example, Imbens & Angrist, 1994). More recent studies, however, provide identification results for the case where the instrument is only valid after conditioning on a set of control variables (Abadie, 2003; Frölich, 2007; Tan, 2006; Uysal, 2011; Donald et al., 2014). In the following, we summarize the assumptions leading to the identification of the LATE when the instrument is valid conditional on the confounders.

A 1 (Exclusion Restriction) $\Pr\left[Y_i(d, z) = Y_i(d, z') \, | X_i\right] = 1$, *for any* $d \in \{0, 1\}$ *and* $z, z' \in$

$\mathcal{Z}$.

Assumption A1 implies that conditional on X_i, the only effect of the instrument on the outcome variable is through the treatment variable. This allows us to use $Y_i(d)$ instead of $Y_i(d, z)$ in 2.1 and 2.2.

A 2 (Unconfounded Type) $\{D_i(z) : z, z' \in \mathcal{Z}\} \perp Z_i | X_i$

Assumption A2 requires that conditional on the confounders, the instrument can be regarded as random. This assumption allows identifying the causal effect of the instrument on the treatment variable.

A 3 (Existence of Compliers) $\Pr[D_i(z) < D_i(z')] > 0$ *for any pair* $z, z' \in \mathcal{Z}$ *with* $z < z'$

Assumption A3 is identical to the first-stage assumption of classical IV estimation. It guarantees that there is a subpopulation whose treatment status is affected by the instrument.

A 4 (Monotonicity) $\Pr[D_i(z) > D_i(z')] = 0$ *for any pair* $z, z' \in \mathcal{Z}$ *with* $z < z'$ *and for all* i *in the set* $\{X = x\}$.

The monotonicity assumption ensures that the instrument moves the treatment only in one direction for all individuals with the same X. Specifically, it excludes the existence of defiers.

A 5 (Overlap) $p(z, x) \equiv \Pr[Z_i = z | X_i = x] < 1$ *for any* $x \in \mathcal{X}$ *and* $z \in \mathcal{Z}$

The overlap assumption implies that no value of X_i perfectly determines Z_i, that is, for any value of $x \in \mathcal{X}$, all values of the instrument Z can be observed. We borrow the terminology from the multivalued treatment variable literature and call the conditional probability $\Pr[Z_i = z | X_i = x]$ the generalized instrument propensity

score (GIPS).

Generalizing Donald et al.'s (2014) notation for a multivalued instrument, we define $W(z) \equiv D(z)Y(1)+(1-D(z))Y(0)$ and $W \equiv W(Z) = \sum_{k=1}^{K} \mathbb{1}\{Z = k\}W(k)$. Thus, as in the binary instrument case, it is easily verifiable that $W = DY(1) + (1 - D)Y(0) = Y$, and

$$\tau_{LATE}(zz') = \frac{\mathrm{E}\left[W_i(z') - W_i(z)\right]}{\mathrm{E}\left[D_i(z') - D_i(z)\right]} \equiv \frac{\tau^W_{z'z}}{\tau^D_{z'z}} \tag{2.3}$$

and

$$\tau_{LATE}(zz'|\mathcal{A}) = \frac{\mathrm{E}\left[W_i(z') - W_i(z)|\, Z_i \in \mathcal{A}\right]}{\mathrm{E}\left[D_i(z') - D_i(z)|\, Z_i \in \mathcal{A}\right]} \equiv \frac{\tau^W_{z'z|\mathcal{A}}}{\tau^D_{z'z|\mathcal{A}}}. \tag{2.4}$$

The unconditional LATE in (2.3), $\tau_{LATE}(zz')$, can be interpreted as the ratio of two treatment effects with a multivalued treatment variable (see Imbens, 2000; Lechner, 2001; Cattaneo, 2010; Uysal, 2015, for a discussion on estimation of the treatment effects with a multivalued treatment variable). The numerator is the average effect on W (equivalently, Y) of instrument z' relative to instrument z, $\tau^W_{z'z}$, whereas the denominator is the average effect on D of instrument z' relative to instrument z, $\tau^D_{z'z}$. Similarly, the conditional LATE in (2.4), $\tau_{LATE}(zz'|\mathcal{A})$, is the ratio of the two average effects of instrument z' relative to z for those who received the instrument with a value in $\mathcal{A}$, that is, the ratio of $\tau^W_{z'z|\mathcal{A}}$ to $\tau^D_{z'z|\mathcal{A}}$.

For both LATE parameters, the numerator and the denominator can be estimated separately. Thus, we will discuss the identification and estimation of $\tau^\eta_{z'z}$ and $\tau^\eta_{z'z|\mathcal{A}}$ for $\eta \in \{W, D\}$. The following identification results follow from the assumptions

stated above. The unconditional mean effect is identified as follows:

$$\begin{aligned}\tau^{\eta}_{z'z} &= \mathrm{E}\left[\mathrm{E}\left[\eta|\, X=x, Z=z'\right] - \mathrm{E}\left[\eta|\, X=x, Z=z\right]\right] \\ &= \int \left(\mathrm{E}\left[\eta|\, X=x, Z=z'\right] - \mathrm{E}\left[\eta|\, X=x, Z=z\right]\right) f_X(x)dx,\end{aligned}$$

where $f_X(x)$ is the density function of X. For the conditional mean effect, we consider the following two cases: (1) the set $\mathcal{A}$ has only one element and (2) the set has two elements. It is straightforward to generalize to sets with more than two elements. Let us start with the identification of $\tau^{\eta}_{z'z|l}$, that is, when $\mathcal{A}=\{l\}$. It is indeed similar to the identification of $\tau^{\eta}_{z'z}$. The difference is that the integration is over the conditional density of X given $Z=l$, $f_{X|Z=l}$, rather than $f_X(x)dx$. Thus,

$$\begin{aligned}\tau^{\eta}_{z'z|l} &= \mathrm{E}\left[\mathrm{E}\left[\eta|\, X=x, Z=z'\right] - \mathrm{E}\left[\eta|\, X=x, Z=z\right]|\, Z=l\right] \qquad (2.5)\\ &= \int \left(\mathrm{E}\left[\eta|\, X=x, Z=z'\right] - \mathrm{E}\left[\eta|\, X=x, Z=z\right]\right) f_{X|Z=l}(x)dx \quad \text{with } l\in\mathcal{Z}\end{aligned}$$

When $\mathcal{A}=\{l,m\}$, we use the total probability law:

$$\tau^{\eta}_{z'z|lm} = \tau^{\eta}_{z'z|l}\Pr\left[Z=l\,|Z\in\{l,m\}\right] + \tau^{\eta}_{z'z|m}\Pr\left[Z=m\,|Z\in\{l,m\}\right], \quad (2.6)$$

where $\tau^{\eta}_{z'z|l} = \mathrm{E}\left[\eta(z')-\eta(z)|\, Z=l\right]$ and $\tau^{\eta}_{z'z|m} = \mathrm{E}\left[\eta(z')-\eta(z)|\, Z=m\right]$. The probabilities in Equation (2.6) are identified based on the data, and the identification of $\tau^{\eta}_{z'z|l}$ and $\tau^{\eta}_{z'z|m}$ is given by (2.5).

The weighting-type identification results can also be written for the treatment effects of interest as follows:

$$\tau^{\eta}_{z'z} = \mathrm{E}\left[\frac{\eta \mathbb{1}\{Z=z'\}}{p(z',x)}\right] - \mathrm{E}\left[\frac{\eta \mathbb{1}\{Z=z\}}{p(z,x)}\right]. \qquad (2.7)$$

As mentioned above, due to the relation in (2.6), the conditional treatment effect $\tau^{\eta}_{z'z|lm}$ can be identified if $\tau^{\eta}_{z'z|l}$ and $\tau^{\eta}_{z'z|m}$ are identified. We can express the conditional treatment effect $\tau^{\eta}_{z'z|l}$ as follows:

$$\tau^{\eta}_{z'z|l} = \frac{\Pr[Z=z']}{\Pr[Z=l]} \mathrm{E}\left[\eta \frac{p(l,x)}{p(z',x)} \middle| Z=z'\right] - \frac{\Pr[Z=z]}{\Pr[Z=l]} \mathrm{E}\left[\eta \frac{p(l,x)}{p(z,x)} \middle| Z=z\right], \quad (2.8)$$

and the weighting identification for $\tau^{\eta}_{z'z|m}$ can be expressed in a similar way.

3 Estimation

We propose estimating the LATE parameters in Equation (2.3) and (2.4) using a weighted regression and a normalized weighting approach. Both methods require estimating the GIPS in the first step. An estimation method suitable for the nature of the instrumental variable can be used.

The weighted regression estimator of $\tau_{LATE}(zz')$ can be derived using the coefficient estimates resulting from the following minimization problem:

$$\min_{\mu_{\eta,k},\alpha_{\eta,k}} \frac{1}{N_k} \sum_{i:Z_i=k} \frac{1}{\hat{p}(k,X_i)} \left(\eta_i - \mu_{\eta,k} - (X_i - \bar{X})'\alpha_{\eta,k}\right)^2 \text{ for } k \in \mathcal{Z},$$

where $\hat{p}(k,X_i)$ is the estimated GIPS, $\mu_{\eta,k}$ is equal to $\mathrm{E}[\mathrm{E}[\eta|\, X=x, Z=k]] = \mathrm{E}[\eta(k)]$, $\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$, and N_k stands for the number of observations with $Z=k$. Therefore, weighted-regression-based estimators of $\tau^{\eta}_{z'z}$ and $\tau_{LATE}(zz')$ are given by

$$\hat{\tau}^{\eta,wreg}_{z'z} = \hat{\mu}_{\eta,z'} - \hat{\mu}_{\eta,z},$$

and

$$\hat{\tau}^{wreg}_{LATE}(zz') = \frac{\hat{\tau}^{W,wreg}_{z'z}}{\hat{\tau}^{D,wreg}_{z'z}},$$

where the superscript $wreg$ stands for the weighted regression method. Note that the

estimator $\hat{\tau}^{wreg}_{LATE}(zz')$ has the so-called double robustness property. Thus, it is consistent if (i) the regression specifications of $\mathrm{E}[\eta(k)]$ for $k = z$ and $k = z'$ are correct, even if the GIPS is misspecified, (ii) the GIPS is correctly specified, but the regression specifications of $\mathrm{E}[\eta(k)]$ for $k = z$ and $k = z'$ are wrong, or (iii) the regression specifications of $\mathrm{E}[\eta(k)]$ for $k = z$ and $k = z'$ and the GIPS are correct.

To estimate the conditional effects, we rewrite the minimization problem as follows, so that we can estimate $\tau^{\eta}_{z'z|l}$ directly using the regression coefficients $\mu_{\eta,k|l}$ for $k = z, z'$:

$$\min_{\mu_{\eta,k|l},\alpha_{\eta,k|l}} \frac{1}{N_l} \sum_{i:Z_i=l} \frac{\hat{r}(l,X_i)}{\hat{r}(k,X_i)} \left(\eta_i - \mu_{\eta,k|l} - (X_i - \bar{X}_l)'\alpha_{\eta,k|l}\right)^2 \text{ for } k \in \mathcal{Z},$$

where $\bar{X}_l = \frac{1}{N_l}\sum_{i:Z_i=l} X_i$. Thus, $\hat{\tau}^{\eta,wreg}_{z'z|l} = \hat{\mu}_{\eta,z'|l} - \hat{\mu}_{\eta,z|l}$, and this can be used together with $\hat{\tau}^{\eta,wreg}_{z'z|m}$ to estimate $\tau^{\eta}_{z'z|lm}$ using (2.6). The unknown probabilities in (2.6) can be estimated by their sample counterparts. Finally, $\hat{\tau}^{\eta}_{z'z|lm}$ for $\eta = W$ and $\eta = D$ is used to estimate the conditional LATE, $\tau_{LATE}(zz'|lm)$.

The weighting estimators of the LATEs are directly linked to the weighting identification results given in (2.7) and (2.8). In fact, one can estimate both treatment effects using the sample analogy principle. However, we propose using weights that add up to one, as in (3.1) and (3.2)

$$\hat{\tau}^{\eta,awe}_{z'z} = \left[\sum_{i=1}^{N} \frac{\eta_i \mathbb{1}\{Z_i = z'\}}{\hat{p}(z',X_i)} \Big/ \sum_{i=1}^{N} \frac{\mathbb{1}\{Z_i = z'\}}{\hat{p}(z',X_i)}\right] - \left[\sum_{i=1}^{N} \frac{\eta_i \mathbb{1}\{Z_i = z\}}{\hat{p}(z,X_i)} \Big/ \sum_{i=1}^{N} \frac{\mathbb{1}\{Z_i = z\}}{\hat{p}(z,X_i)}\right] \tag{3.1}$$

$$\begin{aligned}\hat{\tau}^{\eta,awe}_{z'z|l} = &\left[\sum_{i=1}^{N} \mathbb{1}\{Z_i = z'\}\eta_i \frac{\hat{p}(l,X_i)}{\hat{p}(z',X_i)} \Big/ \sum_{i=1}^{N} \mathbb{1}\{Z_i = z'\}\frac{\hat{p}(l,X_i)}{\hat{p}(z',X_i)}\right] \\ &- \left[\sum_{i=1}^{N} \mathbb{1}\{Z_i = z\}\eta_i \frac{\hat{p}(l,X_i)}{\hat{p}(z,X_i)} \Big/ \sum_{i=1}^{N} \mathbb{1}\{Z_i = z\}\frac{\hat{p}(l,X_i)}{\hat{p}(z,X_i)}\right]\end{aligned} \tag{3.2}$$

The normalized weighting estimator of the ATE has become the standard weighting estimator following Imbens (2004) and Busso et al. (2014). Uysal (2011) suggests using the same normalization to estimate the LATE with a binary instrument. Cattaneo (2010) shows that the normalized estimator for the unconditional treatment effect in the case

of a multivalued treatment variable emerges from the generalized method of moments representation of the treatment effects. However, he does not discuss the estimation of the treatment effect for subpopulations with treatment status equal to certain values. Uysal (2015) also discusses the normalized weights for those subpopulations in case of a multivalued treatment variable. More recently, Sant'Anna & Zhao (2020) have also emphasized the importance of normalization.

For inference, standard errors can be calculated by bootstrap methods or by the asymptotic variance formula. Overall, the asymptotic variance can be derived in a two-step procedure. First, the asymptotic distribution of the numerator and the denominator for each of the LATEs are derived. For the weighted regression method, it is important to take into account that the weights are estimated in the first stage. Then, the asymptotic distribution of the numerator and denominator are used to derive the asymptotic distribution of the ratio estimator. This is a simple application of the Delta method. Alternatively, the entire estimation problem can be written as an M-estimation with moment functions used in different steps. Thus, one can easily derive the asymptotic variance for the LATE of interest by using the general results on M-estimation by Stefanski & Boos (2002). In fact, the normalized weighting estimator corresponds to a weighted regression estimator where the regression models include only a constant. Therefore, the asymptotic distribution of the estimator of the LATE of interest based on adjusted weights can also be derived by using the steps outlined.

4 Data and Institutional Background

Data

We use *Gymnasiasten-Studien*, a longitudinal panel study collecting information on 10th graders attending upper secondary schools (Gymnasien) in the German federal state of North Rhine-Westphalia in 1969. One drawback is that the data are not recent. Two

important aspects encourage us to work with these data despite this drawback. First, the current retention rules are quite similar to those in the 1970s. Second, given that there is not enough empirical evidence on the causal effects of this retention policy for Germany, we still believe the findings of this study will attract attention to and generate more research on the effectiveness of being retained in a grade in the German educational system.

The initial survey consists of 3,295 pupils from 121 classes at 68 upper secondary schools. Approximately 10 years after the original survey, an additional data collection effort took place. The purpose of the additional effort was to collect information on the educational path of the students who participated in the initial survey. In the additional survey, information on the highest school degree achieved, the graduation grades at the end of 10th grade of high school, and, in the case of repetitions, grades leading to the repetition are available, among other school-related information (Wiese et al., 1983). The main information source of the additional survey was the official records of the schools in which the initial survey took place (see Appendix A.1 for the details).

The additional survey is unique in providing such detailed information on school records and crucial for the empirical analysis in this study. Both the outcome and treatment variables are constructed using the additional survey. Our outcome variable of interest is a categorical random variable indicating the highest school degree obtained. The categories are (i) no degree (0), (ii) secondary school leaving certificate (1), (iii) vocational school or technical college degree (2), and (iv) upper (academic) secondary school degree (3).[5] The treatment variable of interest is a binary variable indicating retention ($D = 1$) or promotion ($D = 0$). Furthermore, the multivalued instrumental variable—the number of failing grades in subjects with written exams—and the confounding variables are also constructed using data from this survey. After dropping individuals with missing information in any of these variables and excluding anyone who had been in the 10th grade in the previous school year, we are left with 2,303 students.

[5]In German: 1: "Mittlere Reife, " 2: "Fachabitur, " and 3: "Abitur. "

We also use information from the initial survey to create a set of pre-treatment variables. Pre-treatment variables are important in our analysis, to check the balance between the treated and control subsamples. We construct variables in the following categories: individual characteristics, family background, and cognitive skills. As a measure of cognitive skills, we use the number of correctly solved questions in a standard psychometric Intelligence Structure Test administered in the classroom in the 10th grade. For family background, we use information about the living situation in the household, the father's and mother's education, the mother's employment status, the father's occupational prestige score, and a social class indicator. Finally, we compare gender and age by treatment status.

Descriptive statistics of all these variables are given for the full sample and by treatment status in Table 1. The p-values of the t-test for the difference in means by treatment, as well as normalized differences in means by treatment, are also presented in this table.[6] Very few of the pre-treatment variables differ significantly in their means by treatment assignment, except for average grades by subject at the end of the school year. None of the pre-treatment variables crosses the rule-of-thumb threshold of 0.25 in terms of their normalized differences (Imbens & Wooldridge, 2009; Imbens & Rubin, 2015). On the other hand, the means of the average grades in social science subjects, natural science subjects, and language courses are statistically different from each other in each treatment group.[7] Moreover, the normalized differences for these variables are also larger than the rule-of-thumb threshold.

[6]Following the recommendations of Imbens & Wooldridge (2009) and Imbens & Rubin (2015), we investigate the normalized differences in means. The normalized difference is given by

$$\Delta_X = \frac{\bar{X}_1 - \bar{X}_0}{\sqrt{S_1^2 + S_0^2}},$$

where $\bar{X}_d$ and S_d^2 are the sample mean and sample variance of X, respectively, in the subsample with $D_i = d$, for $d = \{0, 1\}$.

[7]Social science subjects include religion, social studies, German, history, geography, civic studies, and philosophy. Natural science subjects are mathematics, physics, chemistry, and biology. Finally, foreign languages are English, French, Latin, Greek, and Russian.

Table 1: Summary Statistics by Treatment

	Entire Sample	$D = 0$	$D = 1$	p-value	Normalized diff.
Outcome	2.302 (1.00)	2.434 (0.89)	1.336 (1.25)	0.00	0.72
IQ					
1^{st} quintile (lowest)	0.214 (0.41)	0.210 (0.41)	0.238 (0.43)	0.29	0.05
2^{nd} quintile	0.189 (0.39)	0.185 (0.39)	0.217 (0.41)	0.21	0.06
3^{rd} quintile	0.199 (0.40)	0.196 (0.40)	0.224 (0.42)	0.28	0.05
4^{th} quintile	0.206 (0.40)	0.207 (0.41)	0.202 (0.40)	0.86	0.01
5^{th} quintile (highest)	0.188 (0.39)	0.199 (0.40)	0.108 (0.31)	0.00	0.18
Missing	0.003 (0.06)	0.002 (0.05)	0.011 (0.10)	0.03	0.07
Age					
< 15	0.131 (0.34)	0.139 (0.35)	0.072 (0.26)	0.00	0.15
$= 15$	0.501 (0.50)	0.497 (0.50)	0.531 (0.50)	0.29	0.05
$= 16$	0.275 (0.45)	0.270 (0.44)	0.310 (0.46)	0.16	0.06
> 16	0.093 (0.29)	0.094 (0.29)	0.087 (0.28)	0.70	0.02
Female	0.480 (0.50)	0.487 (0.50)	0.430 (0.50)	0.07	0.08
Father's education					
Level 1(lowest)	0.165 (0.37)	0.167 (0.37)	0.152 (0.36)	0.52	0.03
Level 2	0.393 (0.49)	0.391 (0.49)	0.408 (0.49)	0.59	0.02
Level 3	0.193 (0.39)	0.190 (0.39)	0.217 (0.41)	0.29	0.05
Level 4 (highest)	0.228 (0.42)	0.232 (0.42)	0.199 (0.40)	0.21	0.06
Mother's education					
Level 1(lowest)	0.311 (0.46)	0.313 (0.46)	0.296 (0.46)	0.56	0.03
Level 2	0.381 (0.49)	0.374 (0.48)	0.433 (0.50)	0.06	0.09
Level 3	0.177 (0.38)	0.178 (0.38)	0.173 (0.38)	0.86	0.01
Level 4 (highest)	0.115 (0.32)	0.120 (0.33)	0.072 (0.26)	0.02	0.12
Mother's employment					
Currently employed	0.229 (0.42)	0.217 (0.41)	0.318 (0.47)	0.00	0.16
Previously employed	0.468 (0.50)	0.472 (0.50)	0.437 (0.50)	0.27	0.05
Unemployed	0.294 (0.46)	0.302 (0.46)	0.238 (0.43)	0.03	0.10
Missing	0.009 (0.10)	0.009 (0.10)	0.007 (0.08)	0.72	0.02
Father's occupation					
Prestige Score	47.079 (14.67)	47.214 (14.75)	46.094 (14.02)	0.23	0.06
Missing	0.020 (0.14)	0.020 (0.14)	0.018 (0.13)	0.81	0.01
Social Class					
Soc. Class 1 (lowest)	0.214 (0.41)	0.209 (0.41)	0.245 (0.43)	0.17	0.06
Soc. Class 2	0.353 (0.48)	0.352 (0.48)	0.361 (0.48)	0.78	0.01
Soc. Class 3	0.175 (0.38)	0.178 (0.38)	0.159 (0.37)	0.44	0.04
Soc. Class 4	0.162 (0.37)	0.165 (0.37)	0.137 (0.34)	0.23	0.06
Soc. Class 5 (highest)	0.094 (0.29)	0.094 (0.29)	0.094 (0.29)	1.00	0.00
Missing	0.002 (0.04)	0.001 (0.04)	0.004 (0.06)	0.42	0.03
Both parents live in the HH*	0.901 (0.30)	0.906 (0.29)	0.870 (0.34)	0.06	0.08
Number of children in the HH					
1	0.165 (0.37)	0.167 (0.37)	0.148 (0.36)	0.43	0.04
2	0.344 (0.48)	0.336 (0.47)	0.404 (0.49)	0.02	0.10
3	0.251 (0.43)	0.251 (0.43)	0.256 (0.44)	0.84	0.01
4	0.139 (0.35)	0.142 (0.35)	0.119 (0.32)	0.31	0.05
More than 4	0.101 (0.30)	0.105 (0.31)	0.072 (0.26)	0.09	0.08
Average of Social Science Subjects	3.141 (0.66)	3.067 (0.64)	3.684 (0.55)	0.00	0.73
Average of Natural Science Subjects	3.225 (0.66)	3.139 (0.63)	3.856 (0.52)	0.00	0.87
Average of Foreign Languages	3.684 (0.76)	3.556 (0.70)	4.622 (0.48)	0.00	1.26
Previous GPA	3.089 (0.45)	3.042 (0.44)	3.433 (0.29)	0.00	0.74
Number of subjects	11.771 (1.34)	11.793 (1.34)	11.614 (1.37)	0.04	0.09
Number of failing grades	0.66 (0.99)	0.38 (0.53)	2.73 (1.07)	0.00	1.96
N	2,303	2,026	277		

Notes: Authors' calculations. Standard deviations are in parentheses. The p-value is from the t-test for equality of means. Better grades are represented by lower values. HH stands for household. Quartiles and quintiles are calculated using the raw data.

The Decision to Retaining a Student in the Same Grade

In Germany, the regulations for being retained in a grade in grammar schools at the time of the survey were very similar to the current ones. At the end of the school year, students received their final grades for each subject based on their performance during the year, which had been measured by written exams, except in the non-scientific subjects.[8] After these final grades were assigned, a committee of teachers decided whether a student would be promoted to the next grade or would have to repeat the last one. There were official guidelines, but the final decision was left to the discretion of this committee. According to the official guidelines, the retention decision should be based on the student's overall academic performance.[9] A student would usually be promoted to the next grade if s/he had satisfactory grades in all subjects. Retention was recommended in the following situations:

(i) The student has a failing grade (6.0) in one subject with a written exam.

(ii) S/he has an insufficient grade (5.0) in two subjects with written exams.

(iii) S/he has an insufficient grade (5.0) in one subject and only sufficient grades (4.0) in all other scientific subjects.

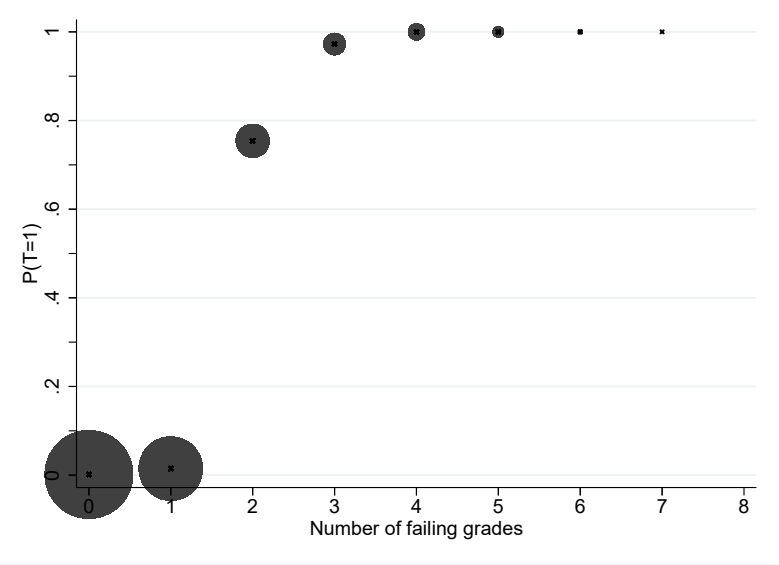
Even though the regulations give the teachers' committee some leeway, it is reasonable to expect that the grades, especially the frequency of receiving failing or insufficient grades,[10] would substantially influence the decision. Indeed, anecdotal evidence provided by retired teachers and individuals who attended school at that time suggests that if a student had failed two subjects, that is, received either 5.0 or 6.0, s/he would be retained. Figure 2 provides further information on the implementation of the grade retention policy. Failing two subjects considerably increased the probability of being retained (to approximately 75%). Students who failed three subjects were almost certainly retained (with a probability of 98%). Thus, the exogenous regulation explained above had a clear effect on the retention probability.

[8] Nonscientific subjects include, for example, physical education, music, and art.

[9] Source: *Amtsblatt des Kultusministeriums*, Land Nordrhein-Westfalen, 1959, 11, pp. 60–61.

[10] In the following, we will only use the term failing grades to refer to grades of both 5.0 and 6.0.

Figure 1: Rate of grade retention



Notes: Each point represents the rate of retention of students having the given number of failing grades (5.0 and 6.0 are the failing grades). The size of each point is proportional to the number of observations. Data source: *Gymnasiasten-Studien*, author's calculations.

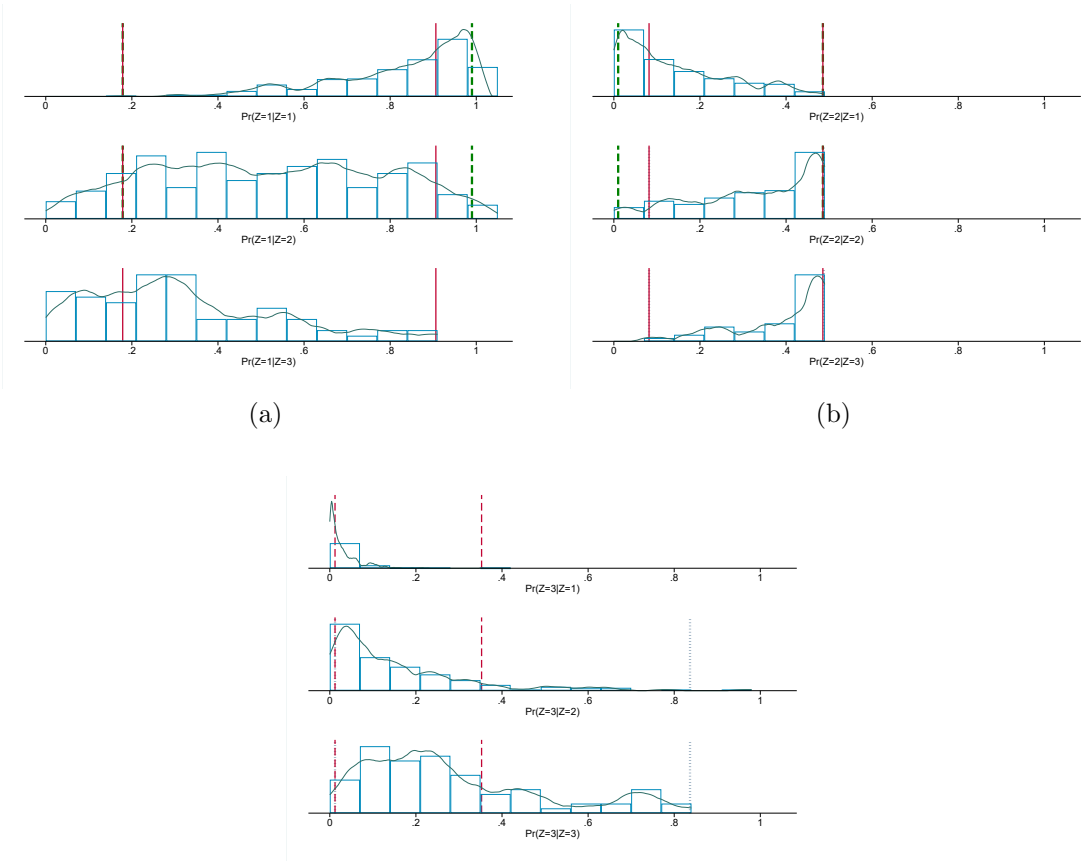
5 Empirical Results

Based on the recommended rules and the discontinuity exhibited due to the rules, we propose to use a student's number of subjects with failing grades as an instrument for grade retention. However, there are two important issues. First, the number of failing grades has an effect on retention only if the student has at least one failing grade and at most three failing grades. Further, the recommended regulations only apply when there is at least one failing grade. Therefore, we argue that the suggested instrument is valid only for students with at least one and at most three failing grades. This idea is similar to regression discontinuity-type identification where the identification of the treatment effect is possible around the discontinuity point. After excluding anyone with no failing grades or more than three failing grades we are left with a sample of 932 students. Second, even though the discontinuity is due to the exogenous recommendations explained above, a

change in the number of failed subjects from one to two, or, similarly, from two to three, is clearly related to overall academic success. Therefore, we argue that only after conditioning on the covariates, particularly on school performance, can the number of failed subjects be considered as a valid instrument. We do not control for the other pre-treatment variables, because they seem to be balanced across the treated and control groups (Table 1). A similar picture arises when we observe the means by treatment, as well as by instrument, after restricting the sample to students with at least one failing grade and at most three failing grades (see Tables A2 and A3, respectively).

Identification of the causal effects of interest hinges upon the validity of the assumption discussed in Section 1. Assumptions A 1 and A 2 together imply that conditional on the school performance measured by the average grades and GPA from the previous year, having one versus two failing grades or having two versus three failing grades can be treated as random. Further, the number of failing grades affects the highest school degree only through grade retention. When the instrument is considered unconditional on the overall school performance measured by average grades, it is very unlikely that it satisfies the exclusion restriction. However, conditioning makes the argument more reliable. Given a student's academic performance, the absolute number of his/her failed subjects is unlikely to affect his/her highest school degree. This is especially true because we restrict our attention to the subsample where the instrument takes only three values, that is, $Z = \{1, 2, 3\}$. Similarly, it is certainly not random whether someone has one or two failing grades, but, conditional on academic performance, it can be considered random. It is clear from Figure 2 that A 3, existence of compliers, is a reasonable assumption. It is also reasonable to assume that the increasing number of failing grades moves the treatment only in one direction (monotonicity). The last assumption, A 5 (Overlap), requires more attention. It is possible that, even after restricting our sample, the overlap assumption is not satisfied. In fact, for some combinations of average grades, it is only possible to have a particular value of Z. Therefore, we first estimate the GIPS using an ordered logit, due to the ordinal nature of the instrumental variable. As explained above, the confounding

Figure 1: Histogram estimates of the GIPS



(a) (b)

(c)

Data source: *Gymnasiasten-Studien* and author's calculations.

variables are the average grades in social science subjects, natural science subjects, and language courses at the end of school year 1969/1970, as well as the average grades from the previous year. We use a linear model of the covariates in levels [11] and visually inspect the histogram estimates of the GIPS by subpopulation. The histogram estimates of the estimated GIPS for each level of the instrument are plotted for individuals with $Z_i = z$ for each $z = 1, 2, 3$ in Figure 1. The histogram estimates indicate some problems with the overlap assumption. Thus, we apply the following rule to determine the common support

[11] We also estimate the GIPS by additionally using the second-order polynomial terms of the covariates. The results do not change qualitatively.

for the estimation of the unconditional treatment effects:

$$\begin{aligned}\mathrm{CS}_z = \Bigg\{ i : \Pr\left[Z_i = z \,|X_i\right] \in \Bigg[\max\left\{ \min_{q \in \mathcal{Z}} \Pr\left[Z_i = z \,|X_i, Z_i = q\right] \right\}, \\ \min\left\{ \max_{q \in \mathcal{Z}} \Pr\left[Z_i = z \,|X_i, Z_i = q\right] \right\} \Bigg] \Bigg\} \\ \mathrm{CS} = \bigcap_{z=1}^{K} \mathrm{CS}_z. \end{aligned} \tag{5.1}$$

Since the estimation of conditional means in the form of $\mathrm{E}\left[\eta(m)|\, l\right]$ with $m \neq l$ needs only the subsamples with $Z \in \mathcal{A} = \{m, l\}$, it requires a less restrictive common support adjustment:

$$\begin{aligned}\mathrm{CS}_{z|\mathcal{A}} = \Bigg\{ i : \Pr\left[Z_i = z \,|X_i\right] \in \Bigg[\max\left\{ \min_{q \in \mathcal{A}} \Pr\left[Z_i = z \,|X_i, Z_i = q\right] \right\}, \\ \min\left\{ \max_{q \in \mathcal{A}} \Pr\left[Z_i = z \,|X_i, Z_i = q\right] \right\} \Bigg] \Bigg\} \\ \mathrm{CS}_{\mathcal{A}} = \bigcap_{z=1}^{K} \mathrm{CS}_{z|\mathcal{A}}. \end{aligned} \tag{5.2}$$

Depending on which unconditional means are required for the LATE of interest, we can define the common support in terms of (5.2). For example, if we are interested in $\tau_{LATE}(12|2)$, then the restriction $\mathrm{CS}_{\mathcal{A}}$ with $\mathcal{A} = \{1, 2\}$ determines the common support. However, if we are interested in $\tau_{LATE}(12|23)$, $\bigcap_{c=1}^{3} \mathrm{CS}_{\mathcal{A}_c}$ with $\mathcal{A}_1 = \{1, 2\}$, $\mathcal{A}_2 = \{1, 3\}$, and $\mathcal{A}_3 = \{2, 3\}$ determines the relevant common support. We illustrate some of the common support rules in Figure 1. The solid red vertical lines indicate the conditions specified by (5.1), whereas the dashed green and dotted blue lines reflect the common support restriction for $\tau_{LATE}(12|12)$ and $\tau_{LATE}(23|23)$, respectively.

Table 2: Estimation Results

	Weighted Regression	Weighting by GIPS		Weighted Regression	Weighting by GIPS	
	Before CS adjustment		nob	after CS adjustment		nob
$\tau_{LATE}(12)$	-0.84	-0.91	921	-0.33	-0.33	565
	(0.35)	(0.75)		(0.20)	(0.20)	
	[-1.10,0.02]	[-1.56,0.02]		[-0.67,-0.02]	[-0.69,-0.04]	
$\tau_{LATE}(12\|1)$	-0.63	-0.95	921	-0.55	-0.95	857
	(0.35)	(1.02)		(0.32)	(0.78)	
	[-0.91,0.20]	[-1.69,0.32]		[-0.83,0.21]	[-1.60,0.35]	
$\tau_{LATE}(12\|2)$	-0.38	-0.53	921	-0.38	-0.53	886
	(0.15)	(0.14)		(0.16)	(0.14)	
	[-0.64,-0.12]	[-0.77,-029]		[-0.64,-0.12]	[-0.77,-0.30]	
$\tau_{LATE}(12\|3)$	-0.30	-1.03	921	-0.18	-0.59	627
	(0.26)	(0.34)		(0.31)	(0.24)	
	[-0.75,0.12]	[-1.44,-0.36]		[-0.79,0.26]	[-1.01,-0.24]	
$\tau_{LATE}(01\|01)$	-0.58	-0.87	921	-0.39	-0.88	821
	(0.29)	(0.58)		(0.24)	(0.55)	
	[-0.82,0.10]	[-0.49,0.15]		[-0.70,0.08]	[-1.45,0.20]	
$\tau_{LATE}(01\|12)$	-0.35	-0.69	921	-0.30	-0.53	627
	(0.17)	(0.18)		(0.20)	(0.16)	
	[-0.63,-0.06]	[-0.94,-0.37]		[-0.65,0.01]	[-0.80,-0.27]	
$\tau_{LATE}(23)$	1.46	3.59	921	-0.04	0.49	565
	(2.01)	(4.93)		(0.98)	(1.46)	
	[-1.79,4.45]	[-2.59,13.17]		[-1.20,2.01]	[-1.56,3.19]	
$\tau_{LATE}(23\|1)$	-0.29	4.26	921	-0.70	4.45	857
	(1.83)	(7.68)		(1.81)	(7.93)	
	[-2.96,2.50]	[-5.23,19.03]		[-3.59,1.68]	[-5.38,19.44]	
$\tau_{LATE}(23\|2)$	0.99	1.18	921	0.99	1.18	921
	(0.88)	(0.88)		(0.88)	(0.89)	
	[-0.21,2.61]	[-0.11,2.71]		[-0.19,2.63]	[-0.11,2.75]	
$\tau_{LATE}(23\|3)$	1.98	5.47	921	1.64	1.71	627
	(1.13)	(5.50)		(1.22)	(1.24)	
	[0.36,4.07]	[0.65,16.73]		[0.35,4.25]	[0.19,4.20]	
$\tau_{LATE}(23\|12)$	-0.02	3.19	921	-0.43	3.62	857
	(1.50)	(4.22)		(1.56)	(4.81)	
	[-2.14,2.37]	[-3.31,10.62]		[-2.89,1.74]	[-3.65,12.09]	
$\tau_{LATE}(23\|23)$	1.26	2.02	921	1.22	1.33	627
	(0.85)	(1.14)		(0.87)	(0.90)	
	[0.04,2.83]	[0.32,3.94]		[0.10,2.94]	[0.10,3.02]	

Notes: Author's calculations. Bootstrapped standard errors and 90% confidence intervals based on 2,000 bootstrap samples are in parentheses and square brackets, respectively. nob stands for number of observations.

We estimate the LATE for two compliers' groups, $\tau_{LATE}(12)$ and $\tau_{LATE}(23)$, as well as the conditional LATEs for all subsamples except for the subsample with $Z \in \{1, 3\}$, using the methods described in Section 2. For the GIPS, we used the average grades in social science subjects, natural science subjects, and language courses at the end of the school year 1969/1970, as well as the average grades from the previous year in levels. The treatment effects are first estimated without applying any common support adjustment.[12]

The unconditional and conditional LATE estimates for the first compliers' groups when the unobserved triplet (D_1, D_2, D_3) is equal to $(0, 1, 1)$ are all negative. The estimated unconditional average effect of retention for this group is -.84 by the weighted regression method and -.91 by the normalized weighting method before adjusting the sample for common support. The common support adjustment leads to a reduction in the negative effect (-0.33) and an increase in the precision of the estimate despite the reduction in the sample size. Thus, grade retention has a significant negative effect on the highest school degree achieved for the individuals who were retained if the instrument value changes from one to two. The conditional LATEs are, in general, similar in magnitude. Note that $\tau_{LATE}(12|23)$ is equivalent to $\tau_{LATT}(12)$. The average effect for the treated compliers is significantly different from zero and negative. The results for the first compliers' group suggest that there is not much heterogeneity within the group.

The estimates for the second group are mostly positive but many of them are not statistically different from zero, for both the unadjusted and adjusted sample. The common support adjustment has a similar effect to the previous treatment effect: the estimates mostly become smaller in magnitude and become more precise. Two conditional LATE estimates, $\tau_{LATE}(23|3)$ and $\tau_{LATE}(23|23)$, although exhibiting very imprecise estimates, do not include zero in their 90% confidence intervals. In fact, more than 95% of the bootstrapped estimates for these effects are positive. For this compliers' group, $\tau_{LATE}(23|3)$ is equal to the LATE

[12]Even though we have not applied common support adjustments yet, we have dropped any observations with predicted probabilities of any instrument level smaller than 10^{-4} to guarantee numerical stability. This procedure decreases the number of observations by 11.

for the treated, $\tau_{LATT}(23)$. The estimates are very large in magnitude, approximately 1.5, with wide confidence intervals. The results for the second group indicate that there is some heterogeneity within this compliers' group. For some in this group, retention has no effect; however, for others, retention has a large positive effect on the highest degree obtained.

Our results suggest that the causal effect of retention is highly heterogeneous. Due to the multivalued instrument, we have two compliers' groups and each of them is affected differently by retention. There are two important implications. First, an ad-hoc abolishment of the retention policy might not be the right strategy given that for some subpopulations, it has positive effects. Second, heterogeneity is worth exploring further, to characterize the groups for which this policy might be harmful or beneficial.

6 Conclusion

In this study, we investigate the identification and estimation of different LATE parameters that are defined in terms of a multivalued instrument. We extend the existing literature on the conditionally valid multivalued instrument by explicitly defining two types of LATEs (conditional and unconditional) and proposing (i) a weighted regression method and (ii) a normalized weighting estimator for both types of LATEs. The proposed weighted regression estimators possess the desired double robustness property.

We use our proposed estimation methods to estimate the causal effects of retention at the end of 10th grade in Germany, based on a rich data set. The data set provides detailed information on grades as well as complete school trajectories, and a rich set of pre-treatment variables such as IQ measures and family background variables. For the identification of the causal effect, we exploit the discontinuity induced by the rules establishing student retention. According to the rules, there are two discontinuities: (i) where the number of failed subjects increases from one to two and (ii) where the number increases from two to three. Therefore, we use the number of failed subjects within this range as our mul-

tivalued instrument, which is assumed to be valid conditional on overall school performance.

We estimate the causal effects for the two compliers' groups. For the first group, whose treatment status changes because the number of failed subjects moves from one to two conditional on overall school performance, the effect of the treatment is negative. The effect in the second group is rather heterogeneous. For the subgroup of students with a smaller number of failed subjects, the effect is statistically not different from zero. However, for the compliers with more failed grades, retention seems to be beneficial. We believe that these findings contribute to the discussion on the effectiveness of grade retention policy in general and, particularly, in Germany.

Conflict of interest

Not applicable.

References

Abadie, A. (2003). Semiparametric instrumental variable estimation of treatment response models. *Journal of Econometrics*, *113*(2), 231–263.

Belloni, A., Chernozhukov, V., Fernández-Val, I., & Hansen, C. (2017). Program evaluation and causal inference with high-dimensional data. *Econometrica*, *85*(1), 233–298.

Busso, M., DiNardo, J., & McCrary, J. (2014). New evidence on the finite sample properties of propensity score reweighting and matching estimators. *The Review of Economics and Statistics*, *96*(5), 885–897.

CAESR (2007). *Dataset Gymnasiastenstudie*. Central Archive for Empirical Social Research, Cologne, Germany.

Calonico, S., Cattaneo, M. D., Farrell, M. H., & Titiunik, R. (2019). Regression discontinuity designs using covariates. *The Review of Economics and Statistics*, *101*(3), 442–451.

Cattaneo, M. D. (2010). Efficient semiparametric estimation of multi-valued treatment effects under ignorability. *Journal of Econometrics*, *155*(2), 138 – 154.

Cattaneo, M. D., Idrobo, N., & Titiunik, R. (2020). *A Practical Introduction to Regression Discontinuity Designs: Foundations*. Elements in Quantitative and Computational Methods for the Social Sciences. Cambridge University Press.

Cockx, B., Picchio, M., & Baert, S. (2019). Modeling the effects of grade retention in high school. *Journal of Applied Econometrics*, *34*(3), 403–424.

de Chaisemartin, C. (2017). Tolerating defiance? local average treatment effects without monotonicity. *Quantitative Economics*, *8*(2), 367–396.

Demski, D., & Liegmann, A. B. (2014). *Facetten von Übergängen im Bildungssystem Nationale und internationale Ergebnisse empirischer Forschung*, chap. Klassenwiederholungen im Kontext von Schul- und Berufsbiographien. Waxmann.

D'Haultfoeuille, X. (2010). A new instrumental method for dealing with endogenous selection. *Journal of Econometrics*, *154*(1), 1–15.

Donald, S. G., Hsu, Y.-C., & Lieli, R. P. (2014). Testing the unconfoundedness assumption via inverse probability weighted estimators of (l)att. *Journal of Business & Economic Statistics*, *32*(3), 395–415.

Dong, Y. (2010). Kept back to get ahead? Kindergarten retention and academic performance. *European Economic Review*, *54*(2), 219–236.

Ehmke, T., Drechsel, B., & Carstensen, C. H. (2008). Grade repetition in PISA-I-Plus: What do students who repeat a class learn in mathematics? *Zeitschrift für Erziehungswissenschaft*, *11*(3), 368–387.

Eide, E. R., & Showalter, M. H. (2001). The effect of grade retention on educational and labor market outcomes. *Economics of Education Review*, *20*(6), 563–576.

Élodie Alet, Bonnal, L., & Favard, P. (2013). Repetition: Medicine for a short-run remission. *Annals of Economics and Statistics*, (111/112), 227–250.

Eren, O., Depew, B., & Barnes, S. (2017). Test-based promotion policies, dropping out, and juvenile crime. *Journal of Public Economics*, *153*, 9–31.

Frölich, M., & Huber, M. (2019). Including covariates in the regression discontinuity design. *Journal of Business & Economic Statistics*, *37*(4), 736–748.

Frölich, M. (2007). Nonparametric IV estimation of local average treatment effects with covariates. *Journal of Econometrics*, *139*(1), 35–75. Endogeneity, instruments and identification.

Fruehwirth, J. C., Navarro, S., & Takahashi, Y. (2016). How the timing of grade retention affects outcomes: Identification and estimation of time-varying treatment effects. *Journal of Labor Economics*, *34*(4), 979–1021.

Gary-Bobo, R. J., Goussé, M., & Robin, J.-M. (2016). Grade retention and unobserved heterogeneity. *Quantitative Economics*, *7*(3), 781–820.

Greene, J. P., & Winters, M. A. (2007). Revisiting grade retention: An evaluation of Florida's test-based promotion policy. *Education Finance and Policy*, *2*(4), 319–340.

Greene, J. P., & Winters, M. A. (2009). The effects of exemptions to Florida's test-based promotion policy: Who is retained?: Who benefits academically? *Economics of Education Review*, *28*(1), 135–142.

Imbens, G., & Rubin, D. (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press.

Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika*, *87*, 706–710.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *The Review of Economics and Statistics*, *86*(1), 4–29.

Imbens, G. W., & Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica*, *62*(2), 467–475.

Imbens, G. W., & Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature*, *47*(1), 5–86.

Jacob, B. A., & Lefgren, L. (2004). Remedial education and student achievement: A regression-discontinuity analysis. *Review of Economics and Statistics*, *86*(1), 226–244.

Jacob, B. A., & Lefgren, L. (2009). The effect of grade retention on high school completion. *American Economic Journal: Applied Economics*, *1*(3), 33–58.

Lechner, M. (2001). Identification and estimation of causal effects of multiple treatments under the conditional independence assumption. In M. Lechner, & F. Pfeiffer (Eds.) *Econometric Evaluation of Labour Market Policies*, (pp. 43–58). Heidelberg: Physica.

Lee, D. S., & Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of Economic Literature*, *48*(2), 281–355.

Manacorda, M. (2012). The cost of grade retention. *Review of Economics and Statistics*, *94*(2), 596–606.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and non-randomized studies. *Journal of Educational Psychology*, *66*, 688–701.

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

Sant'Anna, P. H., & Zhao, J. (2020). Doubly robust difference-in-differences estimators. *Journal of Econometrics*.

Schwerdt, G., West, M. R., & Winters, M. A. (2017). The effects of test-based retention on student outcomes over time: Regression discontinuity evidence from Florida. *Journal of Public Economics*, *152*, 154–169.

Stefanski, L. A., & Boos, D. D. (2002). The calculus of m-estimation. *The American Statistician*, *56*(1), 29–38.

Tan, Z. (2006). Regression and weighting methods for causal inference using instrumental variables. *Journal of the American Statistical Association*, *101*(476), 1607–1618.

Uysal, S. D. (2011). *Three Essays on Doubly Robust Estimation Methods*. Ph.D. thesis, University of Konstanz.

Uysal, S. D. (2015). Doubly robust estimation of causal effects with multivalued treatments: An application to the returns to schooling. *Journal of Applied Econometrics*, *30*(5), 763–786.

Wiese, W., Meulemann, H., & Wieken-Mayser, M. (1983). *Soziale Herkunft und Schullaufbahn : Endbericht über ein Projekt*. Zentralarchiv für Empirische Sozialforschung.

A Appendix

A.1 Data

Among the 68 schools that initially participated in the survey, 6 did not agree to data collection due to privacy concerns. The remaining 62 schools accounted for 91% of the initial sample of students (3,010 students). For approximately 2,400 students, the official records from their initial schools were complete. For less than 2% of the students, it was not possible to follow their school trajectory, either because they were deceased or the records were not readable. For the rest of the students with incomplete records, different strategies were used to complete the information. Parts of the incomplete records belong to students who left their initial school and transferred to another school. The schools to which they were transferred were sent a questionnaire to complete the students' records. For most of the school changers, the new school's name was recorded in the school archives of the initial school. In some cases, however, only the type of the new school was recorded: upper secondary school (Gymnasien), technical high school (Fachoberschulen), college of commerce (Höhere Handelsschulen), or vocational school (Berufsfachschulen). In these cases, all the corresponding types of schools within the residence area were contacted to obtain the information. If these efforts were not successful, but the address of the student with incomplete information was available, then the students were sent a questionnaire to complete the survey. Through these steps, it was possible to collect the information on the school trajectory of approximately 2,700 students. See Table A1 for detailed information on the data collection steps.

A.2 Tables

Table A1: Source of educational path information

Type of data collection		Complete school trajectory collected; Compliance rate		School trajectory not complete	
1) School archives (initial schools)	N	2,402		893	
	%		72.9		27.1
2) Questionnaires sent to the next school attended	N	+131	2,533	-131	762
	%		76.9		23.1
3) Questionnaires sent to students	N	+149	2,682	-149	613
	%		81.4		18.6

Notes: The original table can be found in Wiese et al. (1983) p. 139. Wiese et al. (1983) also provide more detailed information on the data collection efforts.

Table A2: Summary Statistics by Treatment - Restricted Sample

	Restricted Sample	$T = 0$	$T = 1$	p-value	Normalized diff.
Outcome	1.947 (1.09)	2.109 (0.97)	1.412 (1.27)	0.00	0.43
IQ					
1st quintile (lowest)	0.249 (0.43)	0.251 (0.43)	0.241 (0.43)	0.75	0.02
2nd quintile	0.221 (0.42)	0.221 (0.41)	0.222 (0.42)	0.96	0.00
3rd quintile	0.202 (0.40)	0.200 (0.40)	0.208 (0.41)	0.78	0.02
4th quintile	0.194 (0.40)	0.194 (0.40)	0.194 (0.40)	0.99	0.00
5th quintile (highest)	0.129 (0.34)	0.131 (0.34)	0.120 (0.33)	0.68	0.02
Missing	0.005 (0.07)	0.003 (0.05)	0.014 (0.12)	0.05	0.09
Age					
< 15	0.089 (0.28)	0.099 (0.30)	0.056 (0.23)	0.05	0.12
$= 15$	0.446 (0.50)	0.419 (0.49)	0.537 (0.50)	0.00	0.17
$= 16$	0.344 (0.48)	0.349 (0.48)	0.329 (0.47)	0.58	0.03
> 16	0.120 (0.33)	0.133 (0.34)	0.079 (0.27)	0.03	0.12
Female	0.438 (0.50)	0.429 (0.50)	0.468 (0.50)	0.31	0.06
Father's education					
Level 1 (lowest)	0.160 (0.37)	0.166 (0.37)	0.139 (0.35)	0.34	0.05
Level 2	0.381 (0.49)	0.373 (0.48)	0.407 (0.49)	0.36	0.05
Level 3	0.204 (0.40)	0.200 (0.40)	0.218 (0.41)	0.57	0.03
Level 4 (highest)	0.223 (0.42)	0.228 (0.42)	0.208 (0.41)	0.55	0.03
Mother's education					
Level 1 (lowest)	0.289 (0.45)	0.295 (0.46)	0.269 (0.44)	0.46	0.04
Level 2	0.392 (0.49)	0.374 (0.48)	0.449 (0.50)	0.05	0.11
Level 3	0.188 (0.39)	0.190 (0.39)	0.181 (0.39)	0.76	0.02
Level 4 (highest)	0.114 (0.32)	0.126 (0.33)	0.074 (0.26)	0.04	0.12
Mother's employment					
Currently employed	0.245 (0.43)	0.229 (0.42)	0.296 (0.46)	0.04	0.11
Previously employed	0.446 (0.50)	0.440 (0.50)	0.468 (0.50)	0.47	0.04
Unemployed	0.297 (0.46)	0.317 (0.47)	0.231 (0.42)	0.02	0.14
Missing	0.012 (0.11)	0.014 (0.12)	0.005 (0.07)	0.27	0.07
Father's occupation					
Prestige Score	47.130 (14.89)	47.221 (15.11)	46.829 (14.18)	0.73	0.02
Missing	0.023 (0.15)	0.024 (0.15)	0.019 (0.14)	0.65	0.03
Social Class					
Soc. Class 1 (lowest)	0.205 (0.40)	0.196 (0.40)	0.236 (0.43)	0.20	0.07
Soc. Class 2	0.358 (0.48)	0.360 (0.48)	0.352 (0.48)	0.82	0.01
Soc. Class 3	0.166 (0.37)	0.169 (0.38)	0.157 (0.37)	0.69	0.02
Soc. Class 4	0.168 (0.37)	0.172 (0.38)	0.157 (0.37)	0.62	0.03
Soc. Class 5 (highest)	0.101 (0.30)	0.102 (0.30)	0.097 (0.30)	0.84	0.01
Missing	0.001 (0.03)	0.001 (0.04)	0.000 (0.00)	0.58	0.04
Both parents live in the HH*	0.889 (0.31)	0.897 (0.30)	0.866 (0.34)	0.20	0.07
Number of children in the HH					
1	0.170 (0.38)	0.177 (0.38)	0.144 (0.35)	0.25	0.07
2	0.335 (0.47)	0.316 (0.47)	0.398 (0.49)	0.02	0.12
3	0.255 (0.44)	0.246 (0.43)	0.287 (0.45)	0.22	0.07
4	0.139 (0.35)	0.147 (0.35)	0.116 (0.32)	0.25	0.06
More than 4	0.101 (0.30)	0.115 (0.32)	0.056 (0.23)	0.01	0.15
Average of Social Science Subjects	3.385 (0.58)	3.323 (0.57)	3.590 (0.54)	0.00	0.34
Average of Natural Science Subjects	3.547 (0.55)	3.481 (0.54)	3.766 (0.52)	0.00	0.38
Average of Foreign Languages	4.142 (0.55)	4.021 (0.52)	4.542 (0.46)	0.00	0.75
Previous GPA	3.288 (0.34)	3.256 (0.34)	3.393 (0.29)	0.00	0.30
N	932	716	216		

Notes: Author's calculations. Data are restricted to students with at least one failing grade and at most three failing grades. Standard deviations are in parentheses. p-values are from t-tests for equality of means. Lower values represent better grades. *: HH stands for household. Quartiles and quintiles are calculated using the raw data.

Table A3: Summary Statistics by Instrument - Restricted Sample

	Z = 0	Z = 1	Z = 2	Z = 0 vs Z = 1	Z = 1 vs Z = 2	Z = 0 vs Z = 2
Outcome	2.115 (0.98)	1.441 (1.25)	1.630 (1.21)	0.42	0.11	0.31
IQ						
1st quintile (lowest)	0.241 (0.43)	0.251 (0.44)	0.315 (0.47)	0.02	0.10	0.12
2nd quintile	0.218 (0.41)	0.223 (0.42)	0.247 (0.43)	0.01	0.04	0.05
3rd quintile	0.207 (0.41)	0.184 (0.39)	0.192 (0.40)	0.04	0.01	0.03
4th quintile	0.197 (0.40)	0.218 (0.41)	0.110 (0.31)	0.04	0.21	0.17
5th quintile (highest)	0.134 (0.34)	0.112 (0.32)	0.123 (0.33)	0.05	0.03	0.02
Missing	0.003 (0.05)	0.011 (0.11)	0.014 (0.12)	0.07	0.02	0.08
Age						
< 15	0.099 (0.30)	0.056 (0.23)	0.082 (0.28)	0.11	0.07	0.04
$= 15$	0.425 (0.49)	0.486 (0.50)	0.548 (0.50)	0.09	0.09	0.17
$= 16$	0.349 (0.48)	0.335 (0.47)	0.329 (0.47)	0.02	0.01	0.03
> 16	0.128 (0.33)	0.123 (0.33)	0.041 (0.20)	0.01	0.21	0.22
Female	0.426 (0.49)	0.469 (0.50)	0.466 (0.50)	0.06	0.00	0.06
Father's education						
Level 1 (lowest)	0.166 (0.37)	0.123 (0.33)	0.192 (0.40)	0.09	0.13	0.05
Level 2	0.371 (0.48)	0.413 (0.49)	0.397 (0.49)	0.06	0.02	0.04
Level 3	0.196 (0.40)	0.257 (0.44)	0.151 (0.36)	0.10	0.19	0.08
Level 4 (highest)	0.234 (0.42)	0.184 (0.39)	0.219 (0.42)	0.09	0.06	0.02
Mother's education						
Level 1 (lowest)	0.290 (0.45)	0.285 (0.45)	0.288 (0.46)	0.01	0.00	0.00
Level 2	0.375 (0.48)	0.430 (0.50)	0.452 (0.50)	0.08	0.03	0.11
Level 3	0.193 (0.39)	0.184 (0.39)	0.151 (0.36)	0.01	0.06	0.08
Level 4 (highest)	0.128 (0.33)	0.084 (0.28)	0.055 (0.23)	0.10	0.08	0.18
Mother's employment						
Currently employed	0.228 (0.42)	0.296 (0.46)	0.274 (0.45)	0.11	0.03	0.07
Previously employed	0.432 (0.50)	0.508 (0.50)	0.425 (0.50)	0.11	0.12	0.01
Unemployed	0.326 (0.47)	0.190 (0.39)	0.288 (0.46)	0.22	0.16	0.06
Missing	0.013 (0.11)	0.006 (0.07)	0.014 (0.12)	0.06	0.06	0.00
Father's occupation						
Prestige Score	47.100 (15.35)	48.179 (12.98)	44.836 (14.83)	0.05	0.17	0.11
Missing	0.025 (0.16)	0.011 (0.11)	0.027 (0.16)	0.07	0.08	0.01
Social Class						
Soc. Class 1 (lowest)	0.197 (0.40)	0.201 (0.40)	0.288 (0.46)	0.01	0.14	0.15
Soc. Class 2	0.363 (0.48)	0.330 (0.47)	0.384 (0.49)	0.05	0.08	0.03
Soc. Class 3	0.166 (0.37)	0.196 (0.40)	0.096 (0.30)	0.05	0.20	0.15
Soc. Class 4	0.171 (0.38)	0.179 (0.38)	0.123 (0.33)	0.02	0.11	0.09
Soc. Class 5 (highest)	0.101 (0.30)	0.095 (0.29)	0.110 (0.31)	0.02	0.03	0.02
Missing	0.001 (0.04)	0.000 (0.00)	0.000 (0.00)	0.04	.	0.04
Both parents live in the HH*	0.897 (0.30)	0.877 (0.33)	0.849 (0.36)	0.04	0.06	0.10
Number of children in the HH						
1	0.169 (0.38)	0.184 (0.39)	0.137 (0.35)	0.03	0.09	0.06
2	0.322 (0.47)	0.369 (0.48)	0.370 (0.49)	0.07	0.00	0.07
3	0.244 (0.43)	0.251 (0.44)	0.370 (0.49)	0.01	0.18	0.19
4	0.150 (0.36)	0.117 (0.32)	0.096 (0.30)	0.07	0.05	0.12
More than 4	0.115 (0.32)	0.078 (0.27)	0.027 (0.16)	0.09	0.16	0.24
Average of Social Science Subjects	3.326 (0.57)	3.474 (0.56)	3.712 (0.58)	0.19	0.30	0.48
Average of Natural Science Subjects	3.468 (0.54)	3.697 (0.50)	3.916 (0.53)	0.31	0.30	0.59
Average of Foreign Languages	4.001 (0.49)	4.465 (0.55)	4.662 (0.38)	0.63	0.29	1.06
Previous GPA	3.255 (0.34)	3.370 (0.31)	3.391 (0.29)	0.25	0.05	0.30
N	680	179	73			

Notes: Author's calculations. Data are restricted to students with at least one failing grade and at most three failing grades. Standard deviations are in parentheses. p-values are from t-tests for equality of means. Lower values represent better grades. *: HH stands for household. Quartiles and quintiles are calculated using the raw data.