Instrument-based estimation with binarised treatments: issues and tests for the exclusion restriction

by

Andresen, Martin Eckhoff; Huber, Martin

Instrument-based estimation with binarized treatments: Issues and tests for the exclusion restriction

Martin Eckhoff Andresen* and Martin Huber**

*Statistics Norway, **University of Fribourg, Dept. of Economics

Abstract: When estimating local average and marginal treatment effects using instrumental variables (IV), multivalued endogenous treatments are frequently binarized based on a specific threshold in treatment support. Such binarization introduces a violation of the IV exclusion if (i) the IV affects the multivalued treatment within support areas below and/or above the threshold and (ii) such IV-induced changes in the multivalued treatment affect the outcome. We discuss assumptions that satisfy the IV exclusion restriction with a binarized treatment and permit identifying the average effect of (i) the binarized treatment and (ii) unit-level increases in the original multivalued treatment among specific compliers. We derive testable implications of these assumptions and propose tests, which we apply to the estimation of the returns to college graduation instrumented by college proximity.

Keywords: Instrumental variable, LATE, binarized treatment, exclusion restriction.

JEL classification: C12, C21, C26.

We have benefited from comments from John Marshall, participants at EEA-ESEM 2018, seminar participants at the internal research seminars at Statistics Norway and the Department of Economics of the University of Fribourg in Saas-Fee. Andresen acknowledges funding from the Norwegian Research Council, grant no. 237840. Addresses for correspondence: Martin Eckhoff Andresen, Statistics Norway, Akersveien 26, 0177 Oslo, Norway; martin.andresen@ssb.no. Martin Huber, University of Fribourg, Bd. de Pérolles 90, 1700 Fribourg, Switzerland; martin.huber@unifr.ch.

1 Introduction

Instrumental variables (IV) strategies are frequently applied in empirical economics to overcome the endogeneity of a treatment variable, whose causal effect on some outcome variable is of interest to researchers and policy makers. In general, an instrumental variable needs to satisfy relevance and monotonicity conditions, meaning that it monotonically shifts the treatment, as well as validity: The IV must not be associated with treatment-outcome confounders and not directly affect the outcome other than through the treatment, which is known as the IV exclusion restriction. For binary treatment variables, the IV assumptions allow identifying the local average treatment effect (LATE) on the compliers, whose treatment switches as a function of the instrument (Imbens and Angrist, 1994), or the marginal treatment effect (MTE) (Heckman and Vytlacil, 2001, 2005).

For multivalued treatments the instrument identifies a weighted average of effects of unit changes in the treatment on several complier groups defined in terms of treatment-instrument reactions across the support of the treatment. Unfortunately, the size of the effects of unit changes in the treatment are unidentified and the complier groups might be overlapping, see Angrist and Imbens (1995). In practice, multivalued treatments are therefore often binarized based on a specific threshold in the support that appears interesting from a policy perspective, such as whether or not a defendant is incarcerated or not (Bhuller et al., 2020; Loeffler, 2013; Aizer and Doyle, 2015), while the multivalued treatment could perceivably be the length of the prison sentence. As another example, rather than considering years of schooling and aiming at evaluating a weighted average effect of a one year increase in schooling among heterogenous complier groups, one might prefer analyzing a binary indicator for college education for compliers who are induced to finish college by the instrument.

Binarization of treatments are also tempting when analyzing the MTE, i.e. the average effect on those who are indifferent between taking and not taking a binary treatment for a specific level of unobserved resistance to treatment, a framework which requires a binary treatment indicator.

Accordingly, studies estimating MTEs commonly make use of binarized versions of originally multivalued treatments. For instance, Carneiro et al. (2017) evaluate the effects of upper secondary schooling (rather than years of education) using distance to school as instrument. For further examples, see Carneiro et al. (2011); Cornelissen et al. (2018); Felfe and Lalive (2018).

As formally discussed in this paper and also pointed out in Angrist and Imbens (1995), binarizing multivalued treatments generally entails a violation of the IV exclusion restriction.[1] Specifically, the violation occurs if (i) the IV affects the multivalued treatment within support areas below and/or above the threshold for binarization and (ii) such IV-induced changes in the multivalued treatment affect the outcome. In cases where the exclusion restriction holds for the binarized treatment, the identified parameter generally includes the effects of any instrument-induced shifts in the original treatment variable among compliers whose treatment is induced to cross the threshold by the instrument, rather than the effect at the threshold only.

As a methodological contribution, we show that part (i) of the violation of the exclusion restriction has testable implications when the original treatment variable prior to binarization is observed. A necessary (but not sufficient) condition for ruling out 'off-threshold' compliance, i.e. that the IV affects the multivalued treatment within support areas below or above the threshold, is a particular first stage condition. When binarizing the treatment at alternative values across its support, the first stage effect of the instrument must weakly increase up to the threshold chosen by the researcher, and weakly decrease thereafter. This can be tested in a moment inequality framework, see for instance Andrews and Shi (2013). Testing within cells of control variables or even the outcome may improve power, because violations of the first stage conditions in subgroups may be averaged away in the whole sample, as we show in the empirical example.

Furthermore, we consider two special cases of this first stage condition, firstly, that all compliers are situated at the threshold and secondly, that aøø compliers are situated at extreme values of the multivalued treatment. We show that both conditions allow identifying average effects of unit changes in the treatment for a well defined complier group (rather than an average of several

[1] For a related discussion, see Imbens and Rubin (2015), who discuss that the stable unit treatment valuation assumption requires the treatment level not to be coarsened when defining potential outcomes.

heterogeneous complier groups) and that the conditions can be tested by means of standard F-tests. We apply our tests to labor market data from the National Longitudinal Survey of Young Males (NLSYM) as analysed in Card (1995). We consider an indicator for graduating from a 4 year college as our binarized education treatment, where a dummy for proximity to college serves as instrument. Both special cases are soundly rejected, although it should be mentioned that our tests rely on the validity of the instrument for the underlying multivalued treatment. Furthermore, the moment inequality tests suggest that the exclusion restriction might be violated altogether for the binarized treatment.

Marshall (2016) discusses the bias due to binarizing the treatment in the LATE framework, coining the term coarsening bias. Assuming that IV-induced changes in the multivalued treatment affect the outcome only (but not at off-threshold margins), he shows how the instrument identifies the average effect of a unit increase in the multivalued treatment at the threshold among compliers. This allows pinpointing the treatment effect of an instrument-induced shift in treatment from right below to right above the threshold. In contrast, we demonstrate that a causal effect of a binarized treatment is identified even when permitting off-threshold compliers, as long as the threshold captures all compliers in the population. This allows identifying a causal effect under weaker assumptions than Marshall (2016), which, however, includes treatment effects of any off-threshold shifts for threshold-crossing compliers that are induced to increase their treatment by more than one level. In addition, this paper appears to be the first one to propose testing approaches in the context of binarized treatments.

Burgess and Labrecque (2018) point out potential violations of the exclusion restriction when binarizing a multivalued treatment in the context of Mendelian randomization, in which genetic variants are used as instruments. We provide a formal discussion using the potential outcome framework. Even though framed in the IV context, we note that the conditions and methods for testing off-threshold compliance can also be applied in other contexts to check if some variable exclusively affects specific margins of a treatment, conditional on monotonicity and exclusion for the original treatment. For instance, one could test whether a labor market policy only affects the extensive or also the intensive margin of labor supply.

Our paper relates to a growing literature on testing the assumptions for the nonparametric identification of the LATE with binary treatments, that also applies to binarized treatments. Balke and Pearl (1997) derive testable constraints whose violation would imply a negative density of compliers for some value of a binary outcome, even though the lower theoretical bound of densities is zero. Heckman and Vytlacil (2005) generalize these constraints to the continuous outcome case. Kitagawa (2015) proposes a test of the constraints in a moment inequality framework based on resampling variance-weighted Kolmogorov-Smirnov-type statistics on the supremum of violations. Mourifié and Wan (2017) suggest an alternative test that allows controlling for covariates in a user-friendly way.

Huber and Mellace (2015) show that the LATE assumptions imply an alternative set of constraints related to the mean outcomes of non-compliers whose treatment does not react to the instrument. Like most other tests in the literature, this test checks necessary, but not sufficient conditions for instrument validity. That is, the tests are inconsistent in the sense that there may exist data generating processes which satisfy the constraints, but nevertheless violate the LATE assumptions. Sharma (2016) offers an extension by determining the likelihood that the LATE assumptions hold when the testable constraints are satisfied. Specifically, the test defines classes of valid causal models satisfying the LATE assumptions as well as invalid models and compares their marginal likelihood in the observed data. As an alternative strategy, Slichter (2014) suggests testing conditional IV validity by finding covariate values for which the instrument has no first stage and checking whether the instrument is still associated with the dependent variable.

Our tests differ from this and the previously mentioned approaches in that it exploits information in a multivalued treatment prior to binarization, rather than in conditional means or densities of the outcome. We therefore propose a further approach for testing IV validity in cases where the binary treatment was generated from a variable with richer support. Because the tests in the literature generally tests necessary, but not sufficient conditions for instrument validity, failure to reject the null for these tests cannot prove the validity of the exclusion restriction. Our test can thus potentially reject instrument validity in some cases where the test in e.g. Kitagawa (2015) cannot.

In the presence of both a binary and a continuous instrument, Dzemski and Sarnetzki (2014) suggest a nonparametric overidentification tests for IV validity. In contrast, our approach does not require a second IV. Finally, if outcome variables are observed in periods prior to instrument assignment, placebo tests based on estimating the effect of the instrument on pre-instrument outcomes may be performed to check the plausibility of IV validity. Our tests do not rely on the availability of pre-instrument outcomes.

For the multivalued treatment case, Angrist and Imbens (1995) discuss the testable constraint that the cumulative distribution functions of the treatment in the groups receiving and not receiving the instrument must not cross (stochastic dominance), as this would point to a violation of monotonicity, conditional on IV validity. Fiorini and Stevens (2014) point out that testing this necessary condition can also have power against violations of IV validity, conditional on monotonicity. Our framework is different in that we assume that the IV relevance and validity assumptions hold for the original multivalued treatment, but not necessarily for the binarized treatment, for which we test the exclusion restriction. It is important to emphasize how our test utilize other moments of the data than other tests of instrument validity in the literature, and thus may strengthen the plausibility of IV validity even when other tests does not reject.

This paper proceeds as follows. Section 2 introduces the econometric framework, presents minimum assumptions for IV to identify a causal effect of a binarized treatment and discuss the interpretation of this parameter. Section 3 discusses testable implications of the assumptions along with testing approaches. Section 4 presents an application to data from the NLSYM. Section 5 concludes.[2]

[2] Appendix A presents a brief simulation study illustrating how conditioning on the outcome in the tests may increase power.

2 Econometric framework and assumptions

We denote by D a multivalued treatment variable that is ordered discrete, $D \in \{0, 1, ..., J\}$ with $J+1$ being the number of possible treatment doses. An example is years of education. Y denotes the (discrete or continuous) outcome on which the effect ought to be estimated, for instance earnings in the labor market later in life. Under endogeneity, unobserved factors affect both D and Y such that treatment effects cannot be identified from simple comparisons of different levels of the treatment. One possible solution is the availability of an instrumental variable (IV), denoted by Z, which is relevant in the sense that it influences D and valid in the sense that it does not directly affect the outcome and is not associated with unobserved factors influencing the outcome.

For the formal discussion of the identifying assumptions and testable implications, we use the potential outcome framework, see for instance Rubin (1974). Denote by D_z the potential treatment state that would occur if the instrument Z was exogenously set to some value z, and by Y_d the potential outcome with the treatment exogenously set to some value d in the support of D. We will henceforth assume a binary instrument ($Z \in \{1, 0\}$), which simplifies the exposition. but discuss a straightforward extension to a continuous or multivalued instrument at the end of Section 3.

The starting point for our analysis is the standard IV assumptions for heterogeneous treatment effect models, which will be maintained throughout the paper:

Assumption 1 (IV validity and relevance):

(a) $Z \perp (D_1, D_0, Y_0, Y_1, ..., Y_J)$ (IV independence).

(b) $\Pr(D_1 \geq D_0) = 1$ and $\Pr(D_1 > D_0) > 0$ (positive monotonicity).

where "$\perp$" denotes independence. Assumption 1(a) implies two conditions. First, the instrument must be random so that it is unrelated to factors affecting the treatment and/or outcome. Therefore, not only the potential outcomes and treatment states, but also the types, which are defined by the joint potential treatment states, are independent of the instrument. Second, Z must not have a direct effect on Y other than through D, i.e., satisfy an exclusion restriction, which can be seen from the fact that the potential outcomes are only defined in terms of d rather than z and d.[3] The first part of Assumption 1(b) implies that the treatment of any individual does not decrease in the instrument. The second part assumes the existence of individuals whose treatment state positively reacts to the treatment. Both parts together imply a positive first stage effect of the instrument on the treatment: $E(D|Z=1)-E(D|Z=0)>0$. We note that Assumption 1(b) could be replaced by negative monotonicity: $\Pr(D_1 \leq D_0)=1$ and $\Pr(D_1 < D_0)>0$. From an econometric perspective, both versions are equivalent, because when redefining the instrument under negative monotonicity to be $1-Z$, Assumption 1(b) is satisfied.

If D was binary, the local average treatment effect (LATE) on the so-called compliers, who switch treatment from 0 to 1 as a response to a switch in the instrument from 0 to 1, could be identified by the probability limit of two stage least squares (TSLS) or the Wald estimator, see Imbens and Angrist (1994). That is, under Assumption 1 and $D \in \{0,1\}$, $E[Y_1-Y_0|D_1-D_0=1]=\frac{E(Y|Z=1)-E(Y|Z=0)}{E(D|Z=1)-E(D|Z=0)}$. For a multivalued treatment, however, the causal effect for a single complier population defined by specific potential treatment states, e.g. for those increasing treatment from 1 to 2 when the instrument is switched from 0 to 1, is not identified. Angrist and Imbens (1995) show for ordered discrete treatments that it is merely possible to identify a weighted average of causal effects of unit increases in the treatment, $Y_j - Y_{j-1}$, $j \in \{1,...,J\}$. Specifically, the authors show in the proof of their Theorem 1 that under Assumption 1,

$$\frac{E(Y|Z=1)-E(Y|Z=0)}{E(D|Z=1)-E(D|Z=0)}=\sum_{j=1}^{J} w_j \cdot E(Y_j - Y_{j-1}|D_1 \geq j > D_0)=\Delta^w, \tag{1}$$

where the weights are given by

$$w_j = \frac{\Pr(D_1 \geq j > D_0)}{\sum_{j=1}^{J}\Pr(D_1 \geq j > D_0)}. \tag{2}$$

Note that $0 \leq w_j \leq 1$ and $\sum_{j=1}^{J} w_j = 1$. Therefore, the probability limits of TSLS or the Wald

[3] To make these two aspects explicit, Assumption 1(a) may be postulated as two conditions, see Angrist et al. (1996): (i) $Z \perp (D_1, D_0, Y_{1,0}, Y_{0,0}, Y_{1,1}, Y_{0,1}, ..., Y_{1,J}, Y_{0,J})$ and (ii) $Y_{1,d}=Y_{0,d}=Y_d$ for all d in the support of D (exclusion restriction), where $Y_{z,d}$ denotes a potential outcome defined in terms of both the instrument z and the treatment d.

estimator equal a weighted average of effects of unit changes in the treatment on heterogeneous complier groups defined by different margins of the potential treatments. However, the various treatment effects based on unit changes, $E(Y_j - Y_{j-1}|D_1 \geq j > D_0)$, remain themselves unidentified. Furthermore, the complier groups might be overlapping. Some individuals could, for instance, satisfy both $(D_1 \geq j > D_0)$ and $(D_1 \geq j+1 > D_0)$ for some j and therefore be accounted multiple times.

In order to analyze the effects of a particular margin of treatment, many empiricists explicitly or implicitly binarize the multivalued treatment. Examples include the assessment of the effects of a binary indicator for college attendance, instrumented for instance by college proximity (Kane and Rouse, 1993; Carneiro et al., 2011), of fertility measured by a dummy for having three or more children, instrumented by same-sex sibship or twin births (Angrist and Evans, 1998; Black et al., 2005; Mogstad and Wiswall, 2016) and dummies for incarceration, release or disability benefit receipt in the judge leniency literature (Dobbie et al., 2018; Bhuller et al., 2020; Dahl et al., 2014)[4]. Binarization is also common in the literature on the MTE, a parameter that can be regarded as the limit of the LATE for an infinitesimal change in the instrument. See Carneiro et al. (2017, 2011); Cornelissen et al. (2018); Felfe and Lalive (2018) for examples in the context of returns to upper secondary school, college, and child care, respectively.

Let the binarized treatment measure $D_z^* = I\{D_z \geq j^*\}$ denote the potential state of the binarized treatment under $z \in \{0, 1\}$, where $I\{a\}$ is the indicator function that is equal to one when a holds and zero otherwise. j^* denotes a specific threshold value in the support of D. When practitioners analyze these binarized treatments, it is usually not clear what target parameter they have in mind, and the resulting estimate is often interpreted as if the treatment was truly binarized. In contrast, the causal parameter of interest that can be identified under minimal assumptions in this setting can be defined as the average effect among those whose treatment state *passes through* the threshold when switching the instrument from 0 to 1:

[4] Bhuller et al. (2020) provides estimates of the effect of judge stringency on binary dummies for prison sentence exceeding different thresholds in their appendix figure B3. These correspond to the β_j coefficients from this paper. We provide a formal testing framework for instrument validity in this setting.

$$
\begin{aligned}
\Delta^* &= E[Y_{D_1} - Y_{D_0} | D_1^* - D_0^* = 1] = E[Y_{D_1} - Y_{D_0} | D_1 \geq j^* > D_0] \\
&= \sum_{j=1}^{J} E[Y_j - Y_{j-1} | D_1 \geq j > D_0, D_1 \geq j^* > D_0] \cdot \Pr(D_1 \geq j > D_0 | D_1 \geq j^* > D_0),
\end{aligned} \tag{3}
$$

The expression following the second equality in (3) shows that Δ^* is a weighted average of effects among compliers satisfying $D_1^* - D_0^* = 1$, even though they could be defined by different potential (original) treatment states D_0, D_1. That is, the effect refers to all compliers satisfying $D_1 \geq j^* > D_0$, no matter how heterogeneous they are in terms of D_1 and D_0, which is important for interpretation. This differs from the parameter of interest in Marshall (2016), the effect of a unit change in the treatment at the threshold based on switching from $j^* - 1$ to j^* among compliers, which is more challenging to identify. In contrast, our parameter of interest represents a weighted average of all treatment effects for compliers induced to cross j^* by the instrument, not only the treatment effect at the threshold j^*. Rather, it comprises the effects of various treatment shifts from some level below j^* to some level greater than or equal to j^* for all compliers induced to cross j^*. Although practitioners could be interested in the effect of the last unit of treatment that makes an individual cross the threshold, as in Marshall (2016), this is fundamentally harder to identify than Δ^*, which is also analogous to a parameter identified by comparing individuals above and below a treatment cutoff using OLS in the absence of a selection problem.

In the context of returns to college investigated in the empirical application in Section 4, the interpretation of Δ^* is the average causal effect on wages of the extra education obtained by people induced to get college or more by the instrument. This is in contrast to the parameter in Marshall (2016), which is harder to identify, but captures the causal effect of the very last year of college for people induced to finish by the instrument.

Δ^* generally differs from Δ^w identified in (1). The latter identifies an average effect of unit changes. The former corresponds to a total effect, i.e. the sum of effects of unit changes that are weighted with the probability that they occur among compliers crossing the threshold as a response to the instrument. As a matter of fact frequently disregarded by empiricists, Δ^* is generally not identified by the probability limit of the Wald estimator or TSLS based on D^* rather than D,

$$\mathrm{W}^{D^*} = \frac{E(Y|Z=1)-E(Y|Z=0)}{E(D^*|Z=1)-E(D^*|Z=0)}. \tag{4}$$

This is the case despite of the supposed analogy of (4) to the results of Angrist and Imbens (1995) for a (truly) binary treatment. However, a binarization of the treatment variable generally entails a violation of the exclusion restriction such that Assumption 1a for D does not carry over to D^*.

To see this, rewrite the numerator of (4) using the law of total probability and Assumption 1(b) as

$$\begin{aligned}
& E(Y|Z=1)-E(Y|Z=0) \\
= & \sum_{j=1}^{J} E[Y_j - Y_{j-1}|D_1 \geq j > D_0] \cdot \Pr(D_1 \geq j > D_0) \\
= & \sum_{j=1}^{J} E[Y_j - Y_{j-1}|D_1 \geq j > D_0, D_1 \geq j^* > D_0] \cdot \Pr(D_1 \geq j > D_0, D_1 \geq j^* > D_0) \\
+ & \sum_{j=1}^{J} E[Y_j - Y_{j-1}|D_1 \geq j > D_0, I\{D_1 \geq j^* > D_0\}=0] \cdot \Pr(D_1 \geq j > D_0, I\{D_1 \geq j^* > D_0\}=0).
\end{aligned} \tag{5}$$

By summing over j, (5) simplifies to

$$\begin{aligned}
& E(Y|Z=1)-E(Y|Z=0) \\
= & E[Y_{D_1} - Y_{D_0}|D_1 \geq j^* > D_0] \cdot \Pr(D_1 \geq j^* > D_0) \\
+ & E[Y_{D_1} - Y_{D_0}|D_1 > D_0, I\{D_1 \geq j^* > D_0\}=0] \cdot \Pr(D_1 > D_0, I\{D_1 \geq j^* > D_0\}=0).
\end{aligned} \tag{6}$$

Note that the condition $(D_1 > D_0, I\{D_1 \geq j^* > D_0\} = 0)$ captures complier groups whose treatment reacts to the instrument ($D_1 > D_0$), but in a way that it does not cross the threshold

j^* $(I\{D_1 \geq j^* > D_0\} = 0)$. Furthermore, consider the denominator of (4):

$$\begin{aligned} & E(D^*|Z=1) - E(D^*|Z=0) \\ = \;& \Pr(D \geq j^*|Z=1) - \Pr(D \geq j^*|Z=0) = \Pr(D_1 \geq j^*) - \Pr(D_0 \geq j^*) \\ = \;& \Pr(D_1 \geq j^* > D_0) + \Pr(D_0 \geq j^*) - \Pr(D_0 \geq j^*) \\ = \;& \Pr(D_1 \geq j^* > D_0) \qquad (7) \end{aligned}$$

where the second equation follows from Assumption 1(a) and the third from 1(b). Division of (6) by (7) reveals that W^{D^*} does generally not identify Δ^* due to the second line in (6). The latter corresponds to the contribution of compliers whose treatment is not induced to cross j^* by the instrument. For this reason, the parameter of interest Δ^* is only obtained in the special cases that either such off-threshold compliers do not exist or that their average treatment effect is zero, as formalized in Assumptions 2 and 3.

Assumption 2 (zero average treatment effect among non-captured compliers):
$E[Y_{D_1} - Y_{D_0}|D_1 > D_0, I\{D_1 \geq j^* > D_0\} = 0] = 0$.

Assumption 3 (full capturing of compliers by threshold):
$\Pr(D_1 > D_0 \geq j^*) = \Pr(j^* > D_1 > D_0) = 0$.

Assumption 2 postulates the absence of an average causal effect for compliers not captured by the threshold. That is, given a first stage not 'going through' j^*, the average second stage for these compliers must be zero. A related condition has been considered by Marshall (2016) (see his Assumption 5*), requiring that at any treatment level $j \neq j^*$, $E(Y_j|D_1 = j, D_0 = j-1) = E(Y_{j-1}|D_1 = j, D_0 = j-1)$. Note that our Assumption 2 only requires this for (off-threshold) compliers whose multivalued treatment is not induced to cross j^* by the instrument, a considerably weaker condition.

Assumption 3, which can be alternatively formalized as $\Pr(I\{D_1 \geq j^* > D_0\} = 0|D_1 > D_0) = 0$, implies that all compliers are captured by the threshold in the sense that their treatment state is shifted from some $D_0 < j^*$ to some $D_1 \geq j^*$ by the instrument. Thus, there exist no complier groups whose treatment is affected by instrument in a way that D_0, D_1 are either both below or both above the threshold. This rules out first stages not 'going through' the threshold j^*. We note that this assumption is weaker than $\Pr(D_1 \geq j > D_0) = 0$ for $j \neq j^*$, as considered in Marshall (2016) (and imposed in Assumption 4 below): Assumption 3 permits the multivalued treatment to react to the instrument at different margins than (exclusively) j^*, as long as any treatment shifts cross the threshold. Summing up, the IV exclusion restriction fails with binarized treatments if (i) there exist compliers not captured by the definition of D^* and (ii) the instrument-induced changes in treatment affects the outcome of these subjects.

In contrast, if either Assumption 2 or 3 hold,

$$E(Y|Z=1) - E(Y|Z=0) = E[Y_{D_1} - Y_{D_0}|D_1 \geq j^* > D_0] \cdot \Pr(D_1 \geq j^* > D_0), \tag{8}$$

such that $\mathrm{W}^{D^*} = \Delta^*$. Considering the expression after the first equality in (5) reveals that identification is also obtained by combinations of Assumptions 2 and 3 for different subsets of compliers not captured by D^*. For instance, Assumption 3 could hold below the threshold, securing no compliers in this region, while Assumption 2 could hold above the threshold, securing no treatment effects among these compliers.[5] If neither Assumption 2 nor 3 hold, it follows from (6) that the direction of the bias in W^{D^*} is determined by the direction of the average treatment effect among off-threshold compliers. Unfortunately, imposing the popular monotone treatment response (MTR) assumption of Manski and Pepper (2000), which implies that the treatment effect goes in the same direction for both threshold and off-threshold compliers, does not permit bounding the absolute size of Δ^*. On the contrary, MTR implies that W^{D^*} overstates (understates) Δ^* whenever it is positive (negative).

We subsequently discuss two special cases of Assumption 3 for the reason that they allow

[5] A last possibility for identification is the knife-edge case where there exist off-threshold compliers with non-zero effects of the IV-induced changes in treatment, but where these sum to 0, as pointed out by Marshall (2016). Formally, $\sum_{j_0 \neq j^*-1} \sum_{j_1=j_0+1}^{J} E[Y_{j_1} - Y_{j_0}|D_1 = j_1, D_0 = j_0] \Pr(D_1 = j_1, D_0 = j_0) = 0$. This requires treatment effects to go in opposite direction at various levels. We doubt any practitioner would rely on this for identification.

identifying Δ^w, the weighted average effect of unit changes in the treatment, based on W^{D^*}. To this end, rewrite the nominator of (4) as

$$\begin{aligned}
E(Y|Z=1)-E(Y|Z=0) &= \sum_{j=1}^{j^*-1} E[Y_j - Y_{j-1}|D_1 \geq j > D_0] \cdot \Pr(D_1 \geq j > D_0) \\
&+ E[Y_{j^*} - Y_{j^*-1}|D_1 \geq j^* > D_0] \cdot \Pr(D_1 \geq j^* > D_0) \qquad (9) \\
&+ \sum_{j=j^*+1}^{J} E[Y_j - Y_{j-1}|D_1 \geq j > D_0] \cdot \Pr(D_1 \geq j > D_0).
\end{aligned}$$

The first special case occurs if and only if all compliers are concentrated at the threshold such that the instrument has no effect on the treatment at margins of D other than j^*, see also the discussion in Section 3.1 of Angrist and Imbens (1995).

Assumption 4 (concentration of compliers at threshold):

$\sum_{j \neq j^*} \Pr(D_1 \geq j > D_0) = 0$.

It follows from Assumption 4 that (7) and the nominator of (1) are equivalent, implying $E(D|Z = 1) - E(D|Z = 0) = E(D^*|Z = 1) - E(D^*|Z = 0)$ and that $E(Y|Z = 1) - E(Y|Z = 0) = \Delta^* \cdot \Pr(D_1 \geq j^* > D_0)$ in (9) (Angrist and Imbens, 1995). Therefore, $\mathrm{W}^{D^*} = \Delta^* = \Delta^w$, and this coincides with $E(Y_{j^*} - Y_{j^*-1} \mid D_1 = j^*, D_0 = j^* - 1)$, the parameter of interest in Marshall (2016) identified under his Assumptions 2 and 5*. In cases where Assumption 4 is violated, Angrist and Imbens (1995) show that W^{D^*} is larger in absolute terms than Δ^w.

As second special case, assume that all compliers in the population switch their treatment from the lowest ($D_0 = 0$) to the highest ($D_1 = J$) possible treatment value in response to the instrument, while there exist no compliers with other treatment margins affected. This implies that the complier population remains constant across values of j.

Assumption 5 (concentration of compliers at extreme treatment values):

$I\{D_1 \geq j > D_0\} = I\{D_1 \geq j^* > D_0\}$ for all $j, j^* \in \{1, ..., J\}$.

Note that this assumption is stated in terms of indicator functions in contrast to Assumption 4, which is stated in terms of compliance probabilities. The reason is that while constant complier sets across j imply constant compliance probabilities, the converse is not true: There might for example exist compliers that shift D from 0 to 1 and others that shift from 1 to 2 when switching the instrument from 0 to 1. If the shares of these complier groups are the same, the complier probabilities would remain constant across $j \in \{1,2\}$, despite the existence of compliers at intermediate treatment values.

Under Assumption 5, (9) simplifies to

$$\left\{\sum_{j=1}^{J} E[(Y_j - Y_{j-1})|D_1 \geq j^* > D_0]\right\} \cdot \Pr(D_1 \geq j^* > D_0). \tag{10}$$

Therefore, $\mathrm{W}^{D^*} = \Delta^*$ and corresponds to the sum of impacts related to unit changes in treatment D across the entire support. This implies $\Delta^w = \Delta^*/J$, i.e. the average effect of unit changes in the multivalued treatment corresponds to the sum of effects across all possible unit changes divided by the number of possible treatment states J. The reason is that under Assumption 5, the weights in (2) become $\frac{\Pr(D_1 \geq j^* > D_0)}{J \cdot \Pr(D_1 \geq j^* > D_0)} = 1/J$, while in (1), $E(Y_j - Y_{j-1}|D_1 \geq j > D_0) = E(Y_j - Y_{j-1}|D_1 \geq j^* > D_0)$. Therefore,

$$\Delta^w = \frac{E(Y|Z=1) - E(Y|Z=0)}{E(D|Z=1) - E(D|Z=0)} = \frac{E(Y|Z=1) - E(Y|Z=0)}{E(D^*|Z=1) - E(D^*|Z=0)} \Big/ J = \frac{\Delta^*}{J}. \tag{11}$$

3 Testing Assumptions 3, 4, and 5

This section introduces tests for necessary conditions of Assumptions 3, 4, and 5. Notice that all the tests described in this paper rely on the validity of Assumption 1. This means that failure to reject the null hypotheses described could, in principle, be driven either by the invalidity of either Assumption 1 or Assumptions 3, 4 or 5. Importantly, however, we test necessary, not sufficient conditions for validity of Assumptions 3, 4 and 5, and so failure to reject the null can never prove they hold, even when Assumption 1 holds. Rejection of any of the tests, however, points to the invalidity of either Assumption 1 and/or Assumptions 3, 4 or 5.[6]

Under the satisfaction of Assumption 3, it must hold that the share of compliers whose treatment is induced to pass j by the instrument weakly increases when gradually increasing j up to j^*, while weakly decreasing thereafter. The reason is that Assumption 3 requires that j^* captures all compliers, implying that the first stage is maximized at the threshold. Formally, the following weak moment inequality constraints need to hold:

$$\begin{aligned}
&\Pr(D_1 \geq j' > D_0) \geq \Pr(D_1 \geq j'' > D_0) \text{ for all } j^* \geq j' > j'' > 0, \\
&\Pr(D_1 \geq j' > D_0) \leq \Pr(D_1 \geq j'' > D_0) \text{ for all } J \geq j' > j'' \geq j^*.
\end{aligned} \tag{12}$$

Proof. Consider the first line of (12) and note that

$$\begin{aligned}
\Pr(D_1 \geq j' > D_0) &= \Pr(D_1 \geq j' > j'' > D_0) + \Pr(D_1 \geq j' > D_0 \geq j'') \\
&= \Pr(D_1 \geq j'' > D_0) + \Pr(D_1 \geq j' > D_0 \geq j'')
\end{aligned} \tag{13}$$

The first equality follows from the law of total probability and the second from Assumption 3. To see this, note that $\Pr(D_1 \geq j'' > D_0) = \Pr(D_1 \geq j' > j'' > D_0) + \Pr(j' > D_1 \geq j'' > D_0)$. However, by Assumption 3, $\Pr(j' > D_1 \geq j'' > D_0) = 0$ for any $j' \leq j^*$, such that $\Pr(D_1 \geq j'' > D_0) = \Pr(D_1 \geq j' > j'' > D_0)$. Therefore, it follows from $\Pr(D_1 \geq j' > D_0 \geq j'') \geq 0$ that $\Pr(D_1 \geq j' > D_0) \geq \Pr(D_1 \geq j'' > D_0)$. The proof of the second line of (12) is analogous and is therefore omitted. ■

By Assumption 1(a) and (b), (12) implies (in analogy to the discussion in (7) for $\Pr(D_1 \geq j^* > D_0)$) that

$$\begin{aligned}
&\beta_j \geq \beta_{j'} \text{ for all } j^* \geq j > j' > 0, \\
&\beta_j \leq \beta_{j'} \text{ for all } J \geq j > j' \geq j^*,
\end{aligned} \tag{14}$$

[6] One could imagine jointly testing Assumption 1 and Assumptions 3, 4 or 5. We do not pursue this here because rejection of the tests for Assumptions 3, 4 or 5 is more informative because they test necessary, not sufficient conditions.

where $\beta_j = \Pr(D \geq j|Z = 1) - \Pr(D \geq j|Z = 0)$ denotes the first stage effect of Z on the probability that D is larger or equal to some value j. This allows formulating the following null hypothesis for testing Assumption 3, conditional on the satisfaction of Assumption 1:

$$H_0 : \begin{array}{l} \beta_{j+1} - \beta_j \geq 0, \text{ for all } j^* > j > 0, \\ \beta_j - \beta_{j+1} \geq 0, \text{ for all } J > j \geq j^* \end{array} \tag{15}$$

It is important to see that the satisfaction of this null hypothesis is necessary, but not sufficient for Assumption 3. One can easily construct cases in which the weak inequalities hold, even though a subset of individuals complies off threshold. Concerning the practical implementation, it suffices to implement the test for adjacent β_j parameters because of their nested nature: $\beta_2 \geq \beta_0$ provide no additional restrictions on the data when $\beta_2 \geq \beta_1$ and $\beta_1 \geq \beta_0$. These conditions can be verified using testing procedures for moment inequality constraints, see for instance Andrews and Shi (2013).

An implementation is available in the 'cmi_test' command for the statistical software 'Stata' (Andrews et al., 2017), which we use in our application presented in Section 4. We to this end reconsider the first line of (15) and note that

$$\begin{aligned} \beta_{j+1} - \beta_j &= Pr(D \geq j+1 \mid Z=1) - Pr(D \geq j+1 \mid Z=0) \\ &\quad - \Pr(D \geq j \mid Z=1) + \Pr(D \geq j \mid> Z=0) \\ &= \Pr(D = j \mid Z=0) - \Pr(D = j \mid Z=1) \end{aligned} \tag{16}$$

A symmetric argument follows for the second line. Therefore, the sample analog of (15) can be rewritten in the following way based on inverse probability weighting by $E(Z)$ and $1 - E(Z)$:

$$\begin{array}{rll} E(m_j(D,Z)) & \geq 0 & \\ \text{where } m_j(D,Z) & = I\{D = j-1\}\frac{E(Z)-Z}{(1-E(Z))E(Z)} & \text{for } j^* > j > 0 \\ \text{and } m_j(D,Z) & = I\{D = j\}\frac{Z-E(Z)}{(1-E(Z))E(Z)} & \text{for } J > j \geq j^*. \end{array} \tag{17}$$

These constraints match the structure of the 'cmi_test' command of Andrews et al. (2017), which verifies the sample analog of (17). Testing may be implemented both based on Cramer-von-Mises and Kolmogorov-Smirnov-type statistics on average or maximum violations across j, respectively, and both are considered in our empirical application.

Rejection of the test in (15) indicates the presence of non-threshold compliance: Individuals who respond to the instrument, but not in a way that make them cross the threshold j^*. In this case, researchers could look into methods for partial identification with invalid instruments (Flores and Flores-Lagunes, 2013) or sensitivity tests to violations of the exclusion restriction (Huber, 2014). Alternatively, a researcher could rely on Assumption 2 for identification or estimate the the linear IV model using the original treatment D.

Concerning Assumption 4, both a necessary and sufficient condition for its satisfaction, conditional on Assumption 1, is that any first stage effect of Z on the probability that $D \geq j$ must be zero unless $j = j^*$, because all compliers must be located at the threshold. Formally,

$$H_0 : \beta_j = 0 \quad \text{for all } j \neq j^*. \tag{18}$$

Finally, a necessary condition for Assumption 5 is that the first stages or complier probabilities are constant across j. As highlighted in the discussion of Assumption 5 in Section 2, this implies a concentration of compliers at extreme treatment values, but is not sufficient for ruling out other complier groups. Formally, the hypothesis to be tested is

$$H_0 : \beta_j = \beta_{j+1} \quad \text{for all } j < J. \tag{19}$$

Both (18) and (19) can be tested by means of an F-test in a system of equations in which treatment indicator functions $I\{D \geq j\}$ at different values j are regressed on a constant and Z.

If there is heterogeneity in the first stage coefficients across subgroups, performing our tests within cells of X may provide additional power to reject Assumptions 3, 4 or 5. The reason is that violations of e.g. Assumption 3 in some subgroups may be averaged away in the full sample.

Control variables may be included as conditioning set in the moment inequality- and regression-based tests. In (17), for instance, control variables can be considered by replacing $E(Z)$ everywhere with the conditional expectation of Z given the controls, also known as instrument propensity score, and including conditioning on X in the m_j-function, see example 6 in Andrews and Shi (2014). This allows us to jointly test (17) within all cells of X.

Furthermore, the outcome variable may also be used as a conditioning variable in this setup, which may likewise increase power. Although the complier shares in the population cannot be consistently estimated when conditioning on the outcome as the latter is endogenous to the instrument, the sign of any coefficient β_j remains weakly positive when conditioning on Y if monotonicity as postulated in Assumption 1 holds. Therefore, the bias due to conditioning on the outcome cannot entail a violation of the conditions in (15) if Assumption 3 is satisfied. This in turn means that the non-satisfaction of (15) conditional on Y provides evidence for a violation of Assumption 3. The Monte Carlo simulations in Appendix A illustrate this implication and show how conditioning on the outcome can lead to an increase in testing power.

We note that the testing approaches can be extended to multivalued discrete as well as continuous instruments. For multivalued discrete instruments, the conditions given in (15), (18), and (19) must hold when defining $\beta_j = \Pr(D \geq j|Z = z') - \Pr(D \geq j|Z = z'')$ for any values $z' > z''$ in the support of Z. For continuous instruments, the conditions given in (15), (18), and (19) must hold for infinitesimal increases in Z across the entire support of Z. In this case $\beta_j = \frac{\partial \Pr(D \geq j|Z=z)}{\partial z}$ for any z in the support of Z.

Finally, we point out that even though Assumptions 3, 4, and 5 are framed in the context of IV methods, our testing approaches can be applied whenever one is interested in checking if some variable exclusively affects a particular margin of a treatment, conditional on the monotonicity assumption. For instance, a test based on (15) may be used to verify whether a randomized labor market program shifts labor supply only at the extensive margin (working vs. not working), or also at the intensive margin (working more vs. less hours), a test based on (18) may be used to test whether participants are exclusively shifted from no to very low levels of labor supply and a test based on (19) may be used to test whether participants are exclusively shifted from no to full time work.

4 Empirical application

We apply our tests to labor market data previously analysed by Card (1995) that come from the 1966 and 1976 waves of the U.S. National Longitudinal Survey of Young Men (NLSYM). Card (1995) considers a dummy for proximity to a four-year college in 1966 as an instrument for the likely endogenous schooling decision to estimate returns to schooling in 1976. The intuition is that proximity should affect the schooling decision of some individuals, for instance due to costs associated with going to college when not living at home. The original data contain years of schooling as measure of education, but similar to Carneiro et al. (2011), we binarize the treatment to indicate having at least 16 years of education, which roughly corresponds to a four-year college degree.

Table 1: Summary statistics,

Variable	N	mean	s.d.	min	max	comment
Years of schooling	3,010	13.3	2.68	1	18	1976
College dummy	3,010	0.27	0.44	0	1	Dummy for 16 or more years of education
College proximity	3,010	0.68	0.47	0	1	= 1 if near 4-year college in 1966
Log wage	3,010	6.26	0.44	4.6	7.8	log hourly wage in cents, 1976
Age	3,010	28.1	3.14	24	34	
Father's education	2,320	10.0	3.72	0	18	
Mothers' education	2,657	10.3	3.18	0	18	
Region	3,010	4.64	2.27	1	9	Regional dummy, 1966
SMSA	3,010	0.71	0.45	0	1	Metropolitan area of residence dummy
Black	3,010	0.23	0.42	0	1	
Family type	2,796	1.07	0.38	0	2	Single mom / both parents / step-parent

Note: Data source: National Longitudinal Study of Young Men, 1966 and 1976 waves.

The variables used in our analysis are summarized in Table 1. The multivalued treatment is years of schooling in 1976, which varies from 1 to 18 years with a mean of 13.3. Our binarized treatment is a dummy for having 16 or more years of schooling, which has a mean of 0.27. The instrument is a dummy equal to 1 for people living close to a 4-year college in 1966. The outcome is the log of hourly wages in cents, measured in 1976. In addition, we report a range of control variables, including age, parents' education, geographic dummies, race, and a dummy for family type at age 14.

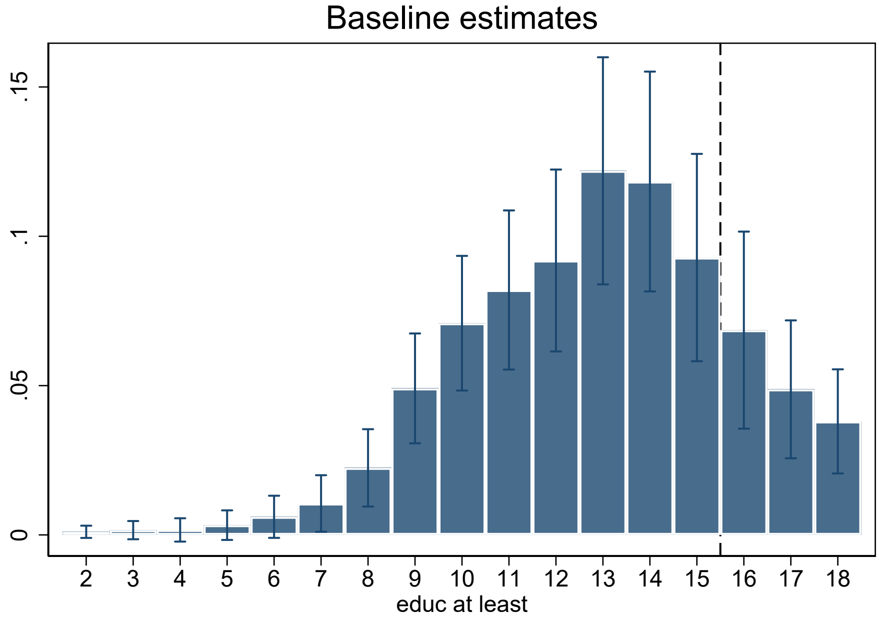

Figure 1: Effects of living close to a four-year college on years of education

Note: Data from NLSYM. Figure shows the estimated impact on binary measures of years of education equal to or above j of living close to a four-year college. The threshold for the binarized treatment is 16 or more years of education as indicated by a dashed line, corresponding roughly to a four-year college degree.

To illustrate our tests, we first estimate the β_j parameters outlined in Section 3, which reflect increases in the probability of having j or more years of schooling when living close to a four-year college compared to living further away, for all margins of education. To this end we estimate a system of equations in which the indicators of having at least j years of education are regressed on the instrument. Figure 1 displays the β_j estimates along with pointwise 95% confidence intervals.

In alternative specifications, we interact the entire specification with fully flexible controls to estimate cell specific β_j coefficients. The reason for this is that proximity to college is likely associated with factors also affecting wages, like local labor market conditions or family background, which would violate Assumption 1. As testing Assumptions 3, 4, and 5 is conditional on Assumption 1, we similarly to Card (1995) control for regional variables (SMSA and region in the US)

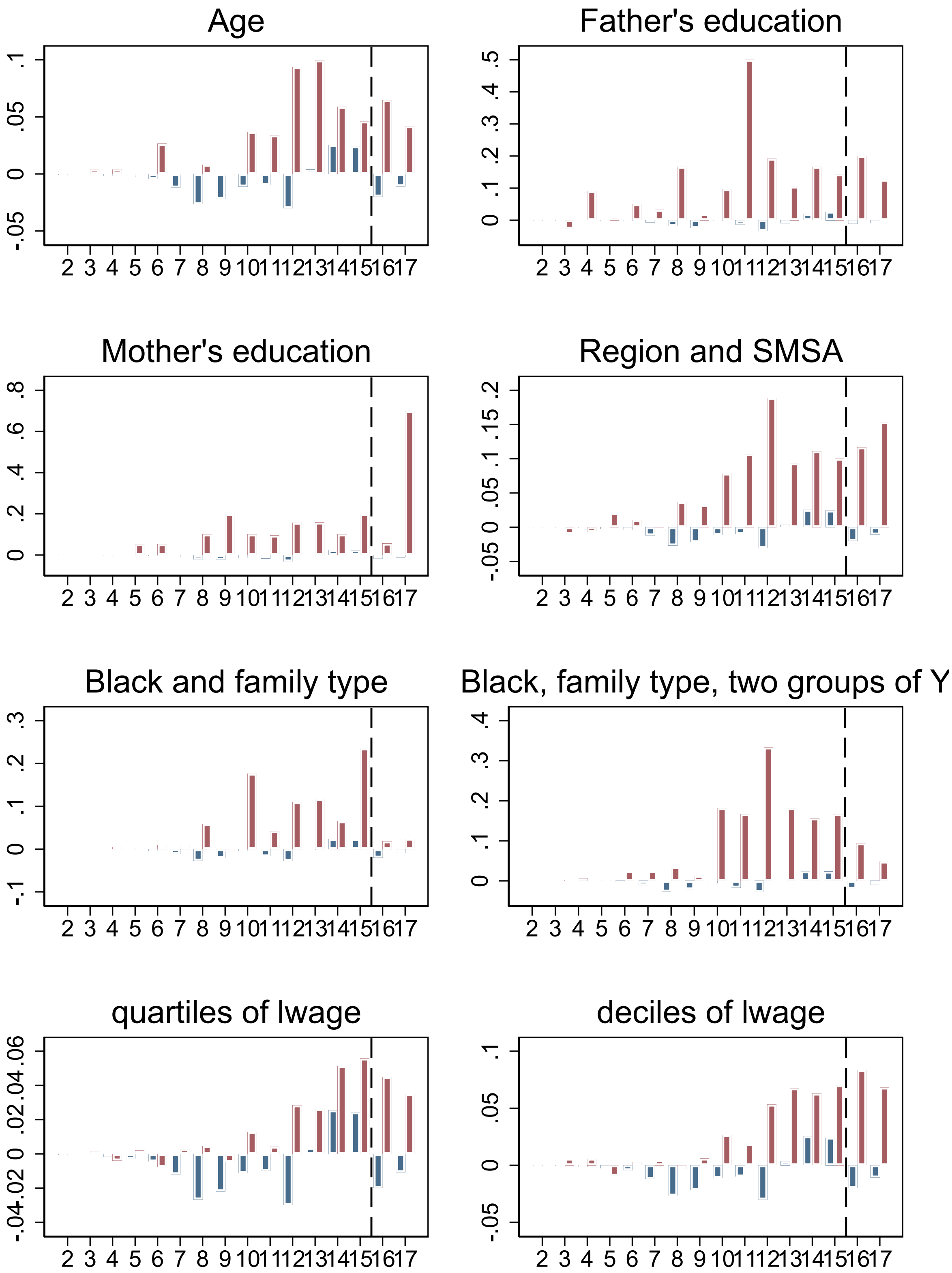

Figure 2: Maximum violations of Assumption 3 across cells

Note: Data from NLSYM. Figure shows the maximum violations across cells of X as indicated in panel title, plotted in red. Violations are $\beta_j - \beta_{j+1}$ for $j < j^*$ and $\beta_{j+1} - \beta_j$ for $j \geq j^*$. For comparison, the violations in the case with no controls are plotted in blue. The threshold for the binarized treatment (j^*) is 16 or more years of education as indicated by a dashed line, corresponding roughly to a four-year college degree.

and socio-economic factors (e.g. parents' education and ethnicity) to increase plausibility of IV exogeneity.

Inspecting Figure 1 allows eye-balling the plausibility of our assumptions for the case with no controls. We observe that the pattern of coefficients are not consistent with Assumption 4, which requires all coefficients except β_{16} to be 0. Neither does it appear to support Assumption 5, which requires the coefficients to be constant across j. Concerning Assumption 3, notice that the dashed line indicating the cutoff value for defining the binarized treatment is to the right of (rather than at) the mode of the β_j estimates, pointing to violations of the conditions in (12).

To formally investigate Assumption 3, we test the constraints in (17) using the 'cmi_test' command of Andrews et al. (2017) based on Cramer-von-Mises and Kolmogorov-Smirnov test statistics.[7] The results are provided in panel A of Table 2. Without including control variables, the p-value of both statistics is 0.049, pointing to a significant violation of the constraints in (15).

When including control variables, we look for violations within cells of X, estimating the β_j coefficients in each cell and testing Assumptions 3, 4 or 5 jointly for all cells. Because there are now multiple sets of β_j coefficients, plotting them all is infeasible. Instead, we plot the maximum violation of Assumption 3 across cells using red bars in Figure 2. For comparison, we also plot the violations from the case with no controls using blue bars. There are indications of violations in some cells at values of j where we found no evidence of violation in the case with no controls. This indicates that there are violations in some groups of X that are averaged away when estimating a single set of β_j coefficients.

The formal tests with controls are found in columns (2)-(9) of Table 2. The violations that were shown graphically in Figure 2 are statistically significant in many of the specifications, indicating violations of Assumption 3 and the presence of off-threshold compliers in at least some cells of X and some off-threshold levels of j. In particular, Assumption 3 is soundly rejected in at least some subgroups of log wage, as seen in columns (7)-(9). When interpreting the results from the tests

[7] A small program for Stata, to be found on github.com/martin-andresen/mvttest, estimates and plots the β_j coefficients, tests Assumption 4 and 5 using F-tests, constructs the moment inequalities and tests them using 'cmi_test'.

Table 2: Tests of instrument validity with a binarized treatment

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
A: Conditional moment inequalities tests of Assumption 3									
Inequalities	16	15	15	13	15	15	15	15	15
Cells of X		11	19	19	18	6	12	4	10
Inequalities tested	16	135	161	162	184	69	117	49	112
Cramer-von-Mises type test statistic									
Test statistic	1.618	2.937	3.483	2.452	4.149	1.961	3.366	3.912	4.623
Critical value 1%	2.186	4.877	4.213	3.981	4.459	3.238	3.798	3.493	4.345
Critical value 5%	1.614	4.202	3.578	3.337	3.827	2.637	3.186	2.794	3.721
Critical value 10%	1.354	3.894	3.258	3.062	3.537	2.367	2.877	2.536	3.422
p-value	0.049	0.459	0.062	0.335	0.024	0.234	0.033	0.002	0.004
Kolmogorov-Smirnov type tests statistic									
Test statistic	13.57	10.73	13.62	11.13	15.00	7.80	20.09	25.09	27.31
Critical value 1%	18.33	21.10	19.78	18.67	20.77	18.56	19.71	18.07	19.65
Critical value 5%	13.53	16.87	15.46	14.62	16.00	14.22	15.27	14.15	15.95
Critical value 10%	11.36	15.12	13.63	12.68	14.26	12.37	13.41	12.31	14.27
p-value	0.049	0.436	0.100	0.191	0.076	0.451	0.009	0.000	0.000
B: F-test of Assumption 4									
F	4.532	1.521	2.003	1.609	2.200	1.871	1.874	1.854	1.340
p-value	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.011
constraints tested	16	135	162	164	184	69	118	49	112
C: F-test of Assumption 5									
F	4.639	1.510	2.082	2.067	2.434	2.074	2.067	2.260	1.631
p-value	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
constraints tested	16	146	180	180	201	74	127	52	120
Controls									
Age		✓							
Fathers' education			✓						
Mothers' education				✓					
Region					✓				
SMSA					✓				
Black						✓	✓		
Family type						✓	✓		
Quantiles of Y							2	4	10
N	3,010	3,010	2,320	2,657	3,010	2,796	2,796	3,010	3,010

Note: Panel A shows test statistics, critical values and resulting p-values from tests of the moment inequalities in (15), tested using cmi_test for Stata (Andrews et al., 2017). Panel B shows the results from an F-test of $\beta_j = 0$ for all $j \neq j^*$ and all cells of X, testing the special case in Assumption 4. Panel C shows F-tests of whether all β_j are the same (within cells of X), testing the special case in Assumption 5. Controls as indicated in the bottom panel. Singleton groups are dropped.

of Assumption 3, it is important to keep in mind that we test necessary, not sufficient conditions for the assumption to hold. Therefore, rejection of the null hypothesis provides evidence against the exclusion restriction, but we can never provide evidence that it holds.

For testing Assumptions 4 and 5, we test the null hypotheses in (18) and (19) using F-tests in our system of equations used to estimate the β_j parameters.[8] The outcomes are displayed in panels B and C of Table 2, respectively. Both assumptions are soundly rejected in all specifications. The results therefore suggest that compliers are not exlusively affected at the threshold, i.e. swithcing from 15 to 16 years of education in response to the instrument, nor exclusively from the lowest to the highest values of education. Therefore, the weighted average of treatment effects based on unit changes, Δ^w, cannot be recovered based on the binarized treatment.

Overall, our results indicate that the exclusion restriction might be violated for the binarized education measure considered. Even though the graphs and estimates suggest that proximity to a four-year college indeed affects education, it may do so not by an exclusive shift towards obtaining at least a four-year college degree. Rather, the instrument seems to also affect the probability of both starting college without finishing and of obtaining a two-year college degree. However, such possibilities are ignored when defining the treatment as a four-year college degree. Judging from the graphs in Figure 1, the exclusion restriction is more likely satisfied if treatment is defined as having at least some college education versus having less education.[9] Yet, we need to bear in mind that even in this case Assumption 3 might be violated, namely if some compliers shift from a two-year college degree to a four-year degree, because we test only necessary, not sufficient conditions for exclusion after binarization.

[8] The system of equations is estimated in a stacked regression using the reghdfe command (Correia, 2014) to account for the covariance of the β_j estimates. Standard errors are clustered at the individual level and robust to heteroskedasticity.

[9] In fact, the constraints in (15) cannot be rejected if the threshold is chosen at the mode of an unimodal set of β_j parameters when there are no control variables. With control variables, however, our tests may indicate violations in some subgroups even in this case.

5 Conclusion

In the context of IV-based estimation, we discussed threats to the exclusion restriction when binarizing a multivalued endogenous treatment. Such a violation occurs whenever (i) the IV affects the multivalued treatment within support areas below and/or above the threshold for binarization and (ii) such IV-induced changes in the multivalued treatment affect the outcome. As a consequence, IV with a binarized treatment identifies the causal effect among individuals whose binary treatment complies with the IV only if either (i) or (ii) can be ruled out. Furthermore, we described the causal parameter that can be identified under these assumptions, which are weaker than previous assumptions in e.g.. Marshall (2016).

More importantly, we showed that (i) has implications that can be tested in a moment inequality framework when the original treatment variable prior to binarization is observed. Furthermore, when ruling out (i) and restricting the support of the multivalued treatment in a particular way, not only the average complier effect of the binarized treatment, but also a weighted average effect of unit changes of the multivalued treatment is recovered. We derived testable implications of these support restrictions that can be verified by standard F-tests. Finally, we provided an empirical illustration to the estimation of returns to a four year college degree, a binarized treatment generated from the multivalued years of education. Our results suggested that the exclusion restriction is violated for such a coarse definition of treatment.

As a final word of caution, we emphasize that the threats to the exclusion restriction not only arise when binarizing a treatment. The issues discussed in this paper prevail whenever the IV affects a finer measure of treatment than used by the researcher in her IV analysis, even when finer treatment measures are not available in the data. Examples include binning a truly continuous treatment into a discrete number of categories or coarsening ordered discrete treatments into a smaller number of categories (e.g. considering low vs. intermediate vs. high levels of education rather than years of schooling). The conditions in this paper highlight under which circumstances the IV validity for the underlying finer treatment measure carries over to a more coarsely defined treatment.

References

Aizer, A. and J. J. Doyle (2015): "Juvenile Incarceration, Human Capital, and Future Crime: Evidence from Randomly Assigned Judges," *The Quarterly Journal of Economics*, 130, 759–803.

Andrews, D. and X. Shi (2013): "Inference Based on Conditional Moment Inequalities," *Econometrica*, 81, 609–666.

Andrews, D. W. and X. Shi (2014): "Nonparametric inference based on conditional moment inequalities," *Journal of Econometrics*, 179, 31 – 45.

Andrews, D. W. K., W. Kim, and X. Shi (2017): "Commands for testing conditional moment inequalities and equalities," *Stata Journal*, 17, 56–72.

Angrist, J. and G. W. Imbens (1995): "Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity," *Journal of American Statistical Association*, 90, 431–442.

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

Angrist, J. D. and W. N. Evans (1998): "Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size," *The American Economic Review*, 88, 450–477.

Balke, A. and J. Pearl (1997): "Bounds on Treatment Effects From Studies With Imperfect Compliance," *Journal of the American Statistical Association*, 92, 1171–1176.

Bhuller, M., G. B. Dahl, K. V. Løken, and M. Mogstad (2020): "Incarceration, Recidivism, and Employment," *Journal of Political Economy*, 128, 1269–1324.

Black, S. E., P. J. Devereux, and K. G. Salvanes (2005): "The More the Merrier? The Effect of Family Size and Birth Order on Children's Education," *The Quarterly Journal of Economics*, 120, 669–700.

Burgess, S. and J. A. Labrecque (2018): "Mendelian randomization with a binary exposure variable: interpretation and presentation of causal estimates," *European Journal of Epidemiology*, 33, 947–952.

Card, D. (1995): "Using Geographic Variation in College Proximity to Estimate the Return to Schooling," in *Aspects of Labor Market Behaviour: Essays in Honour of John Vanderkamp*, ed. by L. Christofides, E. Grant, and R. Swidinsky, Toronto: University of Toronto Press, 201–222.

Carneiro, P., J. J. Heckman, and E. J. Vytlacil (2011): "Estimating Marginal Returns to Education," *American Economic Review*, 101, 2754–81.

Carneiro, P., M. Lokshin, and N. Umapathi (2017): "Average and Marginal Returns to Upper Secondary Schooling in Indonesia," *Journal of Applied Econometrics*, 32, 16–36.

Cornelissen, T., C. Dustmann, A. Raute, and U. Schönberg (2018): "Who Benefits from Universal Child Care? Estimating Marginal Returns to Early Child Care Attendance," *Journal of Political Economy*, 126, 2356–2409.

Correia, S. (2014): "REGHDFE: Stata module to perform linear or instrumental-variable regression absorbing any number of high-dimensional fixed effects," Statistical Software Components, Boston College Department of Economics.

Dahl, G., A. Kostøl, and M. Mogstad (2014): "Family Welfare Cultures," *The Quarterly Journal of Economics*, 129, 1711–1752.

Dobbie, W., J. Goldin, and C. S. Yang (2018): "The Effects of Pretrial Detention on Conviction, Future Crime, and Employment: Evidence from Randomly Assigned Judges," *American Economic Review*, 108, 201–40.

Dzemski, A. and F. Sarnetzki (2014): "Overidentification test in a nonparametric treatment model with unobserved heterogeneity," *mimeo, University of Mannheim.*

Felfe, C. and R. Lalive (2018): "Does early child care affect children's development?" *Journal of Public Economics*, 159, 33 – 53.

Fiorini, M. and K. Stevens (2014): "Monotonicity in IV and fuzzy RD designs - A Guide to Practice," *mimeo, University of Sydney.*

Flores, C. A. and A. Flores-Lagunes (2013): "Partial Identification of Local Average Treatment Effects with an Invalid Instrument," *Journal of Business & Economic Statistics*, 31, 534–545.

Heckman, J. J. and E. Vytlacil (2001): "Local Instrumental Variables," in *Nonlinear Statistical Inference: Essays in Honor of Takeshi Amemiya*, ed. by C. Hsiao, K. Morimune, and J. Powell, Cambridge: Cambridge University Press.

——— (2005): "Structural equations, treatment effects, and econometric policy evaluation 1," *Econometrica*, 73, 669–738.

Huber, M. (2014): "Sensitivity checks for the local average treatment effect," *Economics Letters*, 123, 220–223.

Huber, M. and G. Mellace (2015): "Testing instrument validity for LATE identification based on inequality moment constraints," *Review of Economics and Statistics*, 97, 398–411.

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

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

Kane, T. J. and C. E. Rouse (1993): "Labor Market Returns to Two- and Four-Year Colleges: Is a Credit a Credit and Do Degrees Matter?" Working Paper 4268, National Bureau of Economic Research.

Kitagawa, T. (2015): "A test for instrument validity," *Econometrica*, 83, 2043–2063.

Loeffler, C. E. (2013): "Does Imprisonment Alter the Life Course? Evidence on crime and employment from a natural experiment," *Criminology*, 51, 137–166.

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

Marshall, J. (2016): "Coarsening Bias: How Coarse Treatment Measurement Upwardly Biases Instrumental Variable Estimates," *Political Analysis*, 24, 157–171.

Mogstad, M. and M. Wiswall (2016): "Testing the quantity–quality model of fertility: Estimation using unrestricted family size models," *Quantitative Economics*, 7, 157–192.

Mourifié, I. and Y. Wan (2017): "Testing LATE assumptions," *The Review of Economics and Statistics*, 99, 305–313.

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

Sharma, A. (2016): "Necessary and Probably Sufficient Test for Finding Valid Instrumental Variables," *working paper, Microsoft Research, New York.*

Slichter, D. (2014): "Testing Instrument Validity and Identification with Invalid Instruments," *mimeo, University of Rochester.*

A Appendix: Simulations when conditioning on the outcome

Our simulation study illustrates how conditioning on the outcome may increase the power of testing Assumption 3. The treatment may take three values, $D \in \{0, 1, 2\}$. We set $j^* = 2$ and would like to test whether $\beta_2 - \beta_1 \geq 0$ is violated. The data generating processes (DGP) considered are defined in Table 3. Defiers are ruled out by monotonicity and the population shares of never and always takers are set to 0, too (as they are asymptotically irrelevant for the power of the tests). As the expected value of the test statistic in the full sample corresponds to $E(\beta_2 - \beta_1) = \pi_{12} - \pi_{01} = 0$, we should not be able to detect violations of Assumption 3 in the full sample even thought 30% of the population do not satisfy Assumption 3. However, we may be able to detect such violations in subsamples of Y, because the endogeneity of Y implies that the shares of the different complier groups are different within cells of Y than in the full population. This allows us to detect the presence of the off-threshold complier group C_{01} even if the complier shares are not consistently estimated w.r.t. the total population.

Table 4 shows the results for 1,000 simulations of each of the DGPs outlined in Table 3, using 500 observations per simulation. The first column ("All") provides the results of a test of Assumption 3 based on (15) in the full sample. As expected, we cannot detect violations of Assumption 3 because the presence of the complier group C_{01} is averaged out by the presence of the equally sized complier group C_{12}. Across subsamples of Y in cases (1)-(4), we may detect violations whenever the DGP generates imbalances in the complier groups across cells of Y. While this does not happen in case 1, where compliers violating Assumption 3 are averaged out by non-violating compliers even within cells of Y, we see an increase in testing power in cases (2) - (4), where the different complier groups are shifted differently across Y by the instrument. Figure 3 shows the distribution of the test statistic of each group and each of the four cases, illustrating how conditioning on Y may detect the presence of off-threshold compliers.

Table 3: Data generating process

Complier group	C_{01}	C_{02}	C_{12}									
D_0	0	0	1									
D_1	1	2	2									
$E(Z)$	0.5	0.5	0.5									
Population share π	0.3	0.4	0.3									
	Case 1			Case 2			Case 3			Case 4		
$\Pr(Y = 1 \mid Z = 0, C)$	0.2	0.2	0.2	0.2	0.2	0.2	0.4	0.2	0.1	0.4	0.2	0.1
$\Pr(Y = 1 \mid Z = 1, C)$	0.4	0.4	0.4	0.4	0.5	0.6	0.6	0.4	0.3	0.5	0.4	0.4

Table 4: Simulation results, conditioning on Y

	All	Case 1		Case 2		Case 3		Case 4	
		$Y = 0$	$Y = 1$	$Y = 0$	$Y = 1$	$Y = 0$	$Y = 1$	$Y = 0$	$Y = 1$
β_1	0.70	0.70	0.70	0.65	0.77	0.79	0.57	0.74	0.64
	(0.030)	(0.038)	(0.047)	(0.039)	(0.043)	(0.034)	(0.049)	(0.037)	(0.046)
β_2	0.70	0.70	0.70	0.70	0.70	0.69	0.73	0.69	0.73
	(0.028)	(0.032)	(0.066)	(0.032)	(0.066)	(0.032)	(0.060)	(0.032)	(0.060)
$\beta_2 - \beta_1$	-0.0020	-0.0032	0.00091	-0.053	0.075	0.098	-0.16	0.047	-0.084
	(0.042)	(0.050)	(0.080)	(0.051)	(0.077)	(0.047)	(0.079)	(0.050)	(0.076)
p-value	0.60	0.55		0.39		0.15		0.36	
	(0.37)	(0.31)		(0.31)		(0.20)		(0.30)	
rejection rate	0.032	0.028		0.10		0.42		0.13	
	(0.18)	(0.17)		(0.30)		(0.49)		(0.34)	
true $\beta_2 - \beta_1$	0	0	0	-0.046	0.086	0.134	-0.273	0.090	-0.182

Table reports results from 1,000 simulations of the four different data generating processes described in Table 3, using 500 observations per simulation. The reported values are the means of the respective parameters across the 1000 simulations, standard deviations are reported in parentheses. The rejection rate is based on the 5% level of significance.

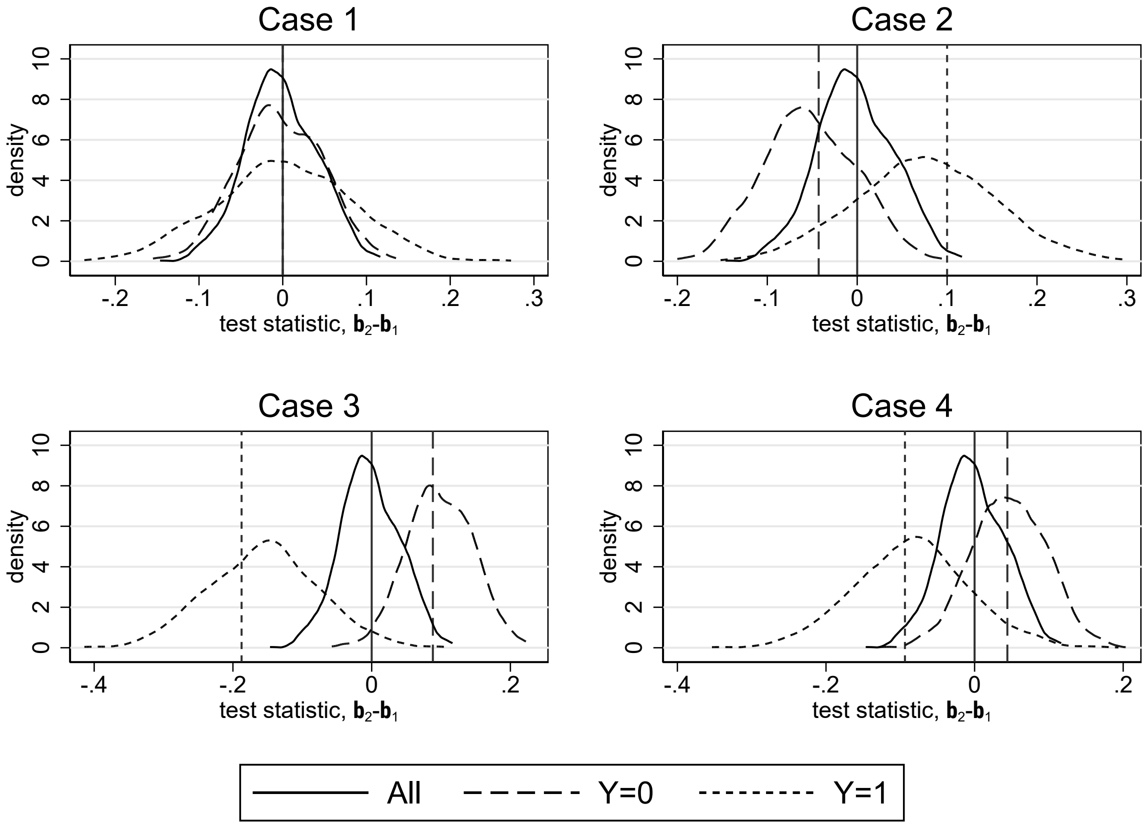

Figure 3: Density plots of test statistics

Note: True values indicated with vertical lines.