The Econometrics Journal

Econometrics Journal (2021), volume **24**, pp. 536–558.
https://doi.org/10.1093/ectj/utab002

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

MARTIN E. ANDRESEN[†] AND MARTIN HUBER[‡]

[†]*Research Department, Statistics Norway, Akersveien 26, 0177 Oslo, Norway.*
Email: martin.eckhoff.andresen@gmail.com

[‡]*Department of Economics, University of Fribourg, Bd. de Pérolles 90, 1700 Fribourg, Switzerland.*
Email: martin.huber@unifr.ch

First version received: 11 May 2020; final version accepted: 9 January 2021.

Summary: When estimating local average and marginal treatment effects using instrumental variables (IVs), multivalued endogenous treatments are frequently converted to binary measures, supposedly to improve interpretability or policy relevance. Such binarisation introduces a violation of the IV exclusion if (a) the IV affects the multivalued treatment within support areas below and/or above the threshold and (b) such IV-induced changes in the multivalued treatment affect the outcome. We discuss assumptions that satisfy the IV exclusion restriction with a binarised treatment and permit identifying the average effect of (a) the binarised treatment and (b) 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*, *binarised treatment*, *exclusion restriction*.

JEL codes: *C12*, *C21*, *C26*.

1. INTRODUCTION

Instrumental variable (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. 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. 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), complicating the interpretation of IV estimates. In practice, multivalued treatments are therefore often binarised based on a specific threshold in the support that

appears interesting from a policy perspective, such as whether or not a defendant is incarcerated (Loeffler, 2013; Aizer and Doyle, 2015; Bhuller et al., 2020), while the multivalued treatment could 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 heterogeneous complier groups, one might prefer analyzing a binary indicator for college education among compliers who are induced to finish college by the instrument.

Binarisation 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 binarised 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 the instrument. For further examples, see Carneiro et al. (2011), Cornelissen et al. (2018), and Felfe and Lalive (2018).

It is well known that such a binarisation may introduce a violation of the IV exclusion restriction, as demonstrated in Angrist and Imbens (1995).[1] Marshall (2016) refers to this issue as coarsening bias and discusses assumptions under which the effect of the multivalued treatment right at the binarisation threshold is identified.

In this paper, we consider a different causal parameter under binarisation, which possibly includes treatment effects away from the threshold and may be identified under weaker assumptions than those in Marshall (2016). Violation of the IV exclusion restriction for this parameter occurs if (a) the IV affects the multivalued treatment within (rather than across) support areas below and/or above the binarisation threshold and (b) such IV-induced changes in the multivalued treatment within support areas affect the outcome. In cases where the exclusion restriction holds for the binarised treatment, the identified parameter includes the effects of any instrument-induced shifts in the multivalued treatment among compliers whose treatment is induced to cross the threshold by the instrument, rather than the effect at the threshold only, as in Marshall (2016).

As a methodological contribution, we show that part (a) of the violation of the exclusion restriction has testable implications when the original treatment variable prior to binarisation 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 binarising 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 (e.g., 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 average out in the whole sample, as we show in the empirical example.

Furthermore, we consider two special cases of this first-stage condition, first, that all compliers are situated at the threshold (as required by Marshall 2016) and, second, that all compliers are situated at the minimum and maximum 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 heterogeneous complier groups) and that the

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

conditions can be tested by means of standard F-tests. Our tests provide a set of diagnostic tools that may guide empiricists when choosing IV specifications based on binarised treatments.[2]

We apply our tests to labour market data from the National Longitudinal Survey of Young Males (NLSYM) as analyzed by Card (1995). We consider an indicator for graduating from a four-year college as our binarised education treatment, where a dummy for proximity to college serves as instrument. Both special cases of the identifying assumptions are soundly rejected. Furthermore, the moment inequality tests suggest that the exclusion restriction might be violated altogether for the binarised treatment (unless one rules out treatment effects among off-threshold compliers).

We build on Angrist and Imbens (1995) who first describe the bias in IV estimates with binarised treatments,[3] and on the identification of the average effect of a unit increase in treatment for compliers at the binarisation threshold provided in Marshall (2016). The latter assumes that IV-induced changes in the multivalued treatment affect the outcome only at the threshold (but not at off-threshold margins). In contrast, we demonstrate that the causal effect of a binarised treatment is identified even when permitting off-threshold compliers, as long as the threshold captures all compliers in the population. This allows for identification under weaker assumptions, but implies that the identified parameter includes treatment effects of any off-threshold shifts for threshold-crossing compliers that are induced to increase their treatment by more than one level.

Our paper contributes to a growing literature on testing the assumptions for the nonparametric identification of the LATE with binary instruments that also applies to binarised treatments. These tests may be based on constraints on the density of compliers (Balke and Pearl, 1997; Heckman and Vytlacil, 2005; Kitagawa, 2015; Mourifié and Wan, 2017), mean outcomes of noncompliers (Huber and Mellace, 2015; Sharma, 2016), reduced form effects for covariate values where there is no first stage (Slichter, 2014), or using additional instruments (Dzemski and Sarnetzki, 2014). Mogstad et al. (2018) study model specification tests in the MTE framework that requires binary treatments. For the multivalued treatment case the instrumental variables assumptions imply the testable condition that the cumulative distribution functions of the treatment conditional on the instrument must not cross across instrument states (Angrist and Imbens, 1995; Fiorini and Stevens, 2014). We appear to be the first to propose formal tests of instrument validity in a setting where the treatment is binarised, including tests for the assumptions in Marshall (2016). Because the tests suggested in the literature generally test necessary, but not sufficient conditions for instrument validity, failure to reject the null does not (even asymptotically) imply the validity of the exclusion restriction. In the context of binarised treatments, this paper adds a further testing approach that may (depending on the data) potentially reject instrument validity in cases where previously suggested methods may not, and be helpful for applied researchers when evaluating potential research designs.

This paper proceeds as follows. Section 2 introduces the econometric framework and presents the baseline assumptions that will be maintained throughout. Section 3 discusses the causal parameters of interest in the literature as well as our alternative parameter along with

[2] 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 verify if some variable exclusively affects specific margins of another variable. One example might be testing whether a (randomised) labour market programme only affects the extensive or also the intensive margin of labour supply.

[3] Burgess and Labrecque (2018) also discuss violations of the exclusion restriction when binarising a multivalued treatment in the context of Mendelian randomisation, in which genetic variants are used as instruments. We provide a formal discussion using the potential outcomes framework.

identifying assumptions. Section 4 presents testable implications of these assumptions and the testing approaches. Section 5 presents an application to data from the NLSYM. Section 6 concludes.[4]

2. ECONOMETRIC FRAMEWORK AND BASELINE ASSUMPTIONS

We indicate with 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 labour 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 (e.g., Rubin, 1974). D_z denotes 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 4.

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.[5] 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.

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

[5] To make these two aspects explicit, Assumption 1(a) may be postulated as two conditions, see Angrist et al. (1996): (a) $Z \perp (D_1, D_0, Y_{1,0}, Y_{0,0}, Y_{1,1}, Y_{0,1}, ..., Y_{1,J}, Y_{0,J})$ and (b) $Y_{1,d} = Y_{0,d} = Y_d$ 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.

In order to analyze the effects of a particular margin of treatment, many empiricists explicitly or implicitly binarise 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; Card, 1995; Carneiro et al., 2011), 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; Dahl et al., 2014; Bhuller et al., 2020).[6] Binarisation 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. (2011, 2017), Cornelissen et al. (2018), and Felfe and Lalive (2018) for examples in the context of returns to upper secondary school, college, and child care, respectively.

Let the binarised treatment measure $D_z^* = I\{D_z \geq j^*\}$ denote the potential state of the binarised treatment under $z \in \{0, 1\}$, where $I\{a\}$ is the indicator function that is equal to one when a holds and zero otherwise. And $j^* > 0$ denotes a specific threshold value in the support of D.

3. PARAMETERS OF INTEREST AND IDENTIFYING ASSUMPTIONS

If D was binary, the local average treatment effect (LATE) on the so-called compliers, which 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 $D \in \{0, 1\}$,

$$W^D = \frac{E(Y|Z=1) - E(Y|Z=0)}{E(D|Z=1) - E(D|Z=0)} = E[Y_1 - Y_0|D_1 - D_0 = 1] = \Delta. \tag{3.1}$$

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,

$$\frac{E(Y|Z=1) - E(Y|Z=0)}{E(D|Z=1) - E(D|Z=0)} = \sum_{j=1}^{J} w_j \cdot E(Y_j - Y_{j-1}|D_1 \geq j > D_0) = \Delta^w, \tag{3.2}$$

where the weights are given by

$$w_j = \frac{\Pr(D_1 \geq j > D_0)}{\sum_{j=1}^{J} \Pr(D_1 \geq j > D_0)}. \tag{3.3}$$

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 estimator equal a weighted average of effects of unit changes in the treatment on

[6] 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.

heterogeneous complier groups defined by different margins of the potential treatments. However, the average treatment effects of unit changes for compliers, $E(Y_j - Y_{j-1}|D_1 \geq j > D_0)$, 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, complicating the interpretation of Δ^w.

When practitioners analyze binarised treatments, it is not always clear what target parameter they are aiming at, and results are often interpreted as if the treatment was truly binary. To avoid such confusion and misinterpretation, we are explicit about the causal parameter that we consider (and the required identifying assumptions), which corresponds to the average effect of the (multivalued) treatment among those induced to *cross the threshold* j^* in response to the instrument:

$$\begin{aligned}\Delta^* &= E[Y_{D_1} - Y_{D_0}|D_1^* - D_0^* = 1] = E[Y_{D_1} - Y_{D_0}|D_1 \geq j^* > D_0] \\ &= \sum_{j=1}^{J} E[Y_j - Y_{j-1}|D_1 \geq j > D_0, D_1 \geq j^* > D_0] \cdot \Pr(D_1 \geq j > D_0|D_1 \geq j^* > D_0).\end{aligned} \tag{3.4}$$

The expression following the second equality in (3.4) 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 $D_1 \geq j^* > D_0$, regardless of how heterogeneous they are in terms of D_1 and D_0, which is important for interpretation. Thus, the parameter Δ^* contains the sum of effects of treatment for individuals induced to cross the threshold j^* by the instrument. This is in contrast to Δ^w, which is an average effect of unit changes for all individuals affected by the instrument, not only those crossing the threshold.

In the context of the returns to college investigated in the empirical application in Section 5, Δ^* is the average causal effect on wages of the extra education obtained by individuals that have been induced by the instrument to attend college for at least four years. Even though this parameter generally averages over several education levels depending how many years of schooling the individuals would have achieved with and without the instrument, it may be relevant for assessing the average effect of policies aimed at increasing access to college, as we discuss later in this section.

Unfortunately, Δ^* is generally not identified by the probability limit of the Wald estimator or TSLS based on D^* rather than D under Assumption 1 alone,

$$\mathrm{W}^{D^*} = \frac{E(Y|Z=1) - E(Y|Z=0)}{E(D^*|Z=1) - E(D^*|Z=0)}, \tag{3.5}$$

despite the supposed analogy of (3.5) to the results of Angrist and Imbens (1995) for a (truly) binary treatment. This is because a binarisation of the treatment variable generally entails a violation of the exclusion restriction, such that Assumption 1(a) for D does not carry over to D^*. To see this, rewrite the numerator of (3.5) using the law of total probability and Assumption 1(b)

as

$$
\begin{aligned}
&E(Y|Z=1)-E(Y|Z=0)\\
&=\sum_{j=1}^{J} E[Y_j - Y_{j-1}|D_1 \geq j > D_0]\cdot \Pr(D_1 \geq j > D_0)\\
&=\sum_{j=1}^{J} E[Y_j - Y_{j-1}|D_1 \geq j > D_0, D_1 \geq j^* > D_0]\cdot \Pr(D_1 \geq j > D_0, D_1 \geq j^* > D_0)\\
&\quad+\sum_{j=1}^{J} E[Y_j - Y_{j-1}|D_1 \geq j > D_0, I\{D_1 \geq j^* > D_0\}=0]\\
&\quad\cdot \Pr(D_1 \geq j > D_0, I\{D_1 \geq j^* > D_0\}=0). \qquad (3.6)
\end{aligned}
$$

By summing over j, (3.6) simplifies to

$$
\begin{aligned}
&E(Y|Z=1)-E(Y|Z=0)\\
&=E[Y_{D_1} - Y_{D_0}|D_1 \geq j^* > D_0]\cdot \Pr(D_1 \geq j^* > D_0) \qquad (3.7)\\
&\quad+E[Y_{D_1} - Y_{D_0}|D_1 > D_0, I\{D_1 \geq j^* > D_0\}=0]\cdot \Pr(D_1 > D_0, I\{D_1 \geq j^* > D_0\}=0).
\end{aligned}
$$

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^*. Furthermore, consider the denominator of (3.5):

$$
\begin{aligned}
&E(D^*|Z=1)-E(D^*|Z=0)\\
&=\Pr(D \geq j^*|Z=1)-\Pr(D \geq j^*|Z=0)=\Pr(D_1 \geq j^*)-\Pr(D_0 \geq j^*)\\
&=\Pr(D_1 \geq j^* > D_0)+\Pr(D_0 \geq j^*)-\Pr(D_0 \geq j^*)\\
&=\Pr(D_1 \geq j^* > D_0), \qquad (3.8)
\end{aligned}
$$

where the second equation follows from Assumption 1(a) and the third from 1(b). Division of (3.7) by (3.8) reveals that W^{D^*} does generally not identify Δ^* due to the second line in (3.7). The latter contains the effects of the instrument on compliers whose treatment is not induced to cross j^* by the instrument. For this reason, the parameter of interest Δ^* is only obtained in special cases where either such off-threshold compliers do not exist or where their average treatment effect is zero, as formalised in Assumptions 2 and 3.

ASSUMPTION 2 (ZERO AVERAGE TREATMENT EFFECT AMONG NONCAPTURED 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.

Assumption 3, which can be alternatively formalised 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 the 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^*. Summing up, the IV exclusion restriction fails with binarised treatments if (a) there exist compliers not captured by the definition of D^* and (b) the instrument-induced changes in treatment affects the outcome of these subjects. In contrast, Δ^* is identified under either Assumption 2 or 3 as postulated in Proposition 1.

PROPOSITION 1 (IDENTIFICATION OF Δ^* UNDER ASSUMPTION *2* OR *3*)*: Under Assumption 1 and either Assumption 2 or 3,*

$$E(Y|Z=1) - E(Y|Z=0) = E[Y_{D_1} - Y_{D_0}|D_1 \geq j^* > D_0] \cdot \Pr(D_1 \geq j^* > D_0), \quad (3.9)$$

such that $W^{D^*} = \Delta^*$.

Considering the expression after the first equality in (3.6) 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.[7]

If neither Assumption 2 nor 3 holds, it follows from (3.7) 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 has the same sign 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).

Applied researchers often implicitly or explicitly impose the standard LATE assumptions directly on D^* and proceed without discussing these issues. The above results make clear what this implies in a setting where the treatment is binarised (or the underlying treatment could be multivalued, even if unobserved), and two comments are in order. First, imposing the standard LATE assumptions directly on D^* implies imposing either Assumption 2 or Assumption 3 for all off-threshold complier groups. This should be discussed explicitly when evaluating the plausibility of the identifying assumptions in such a setting. Second, the identified treatment effect will include effects of nonthreshold shifts in treatment caused by the instrument, but only for compliers also induced to cross the threshold. This complicates the interpretation when compared to the LATE with a truly binary treatment.

Marshall (2016) discusses the problem of binarisation and proposes an alternative parameter of interest:

$$\Delta^M = E[Y_{j^*} - Y_{j^*-1}|D_1 = j^*, D_0 = j^* - 1], \quad (3.10)$$

which is the average effect of the shift from right below to right above the threshold for compliers at this margin. As shown by Marshall (2016), identification of this parameter requires one of the following stronger versions of the assumptions above to hold in addition to Assumption 1:

[7] A last possibility for identification is the knife-edge case where there exist off-threshold compliers with nonzero effects of the IV-induced changes in treatment, but where these sum to 0, as pointed out by Marshall (2016) for his alternative parameter of interest discussed further below. 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 directions at various levels, which appears to be an unattractive assumption from a practical perspective.

ASSUMPTION 2* (ZERO AVERAGE TREATMENT EFFECT FOR OFF-THRESHOLD COMPLIERS): $E[Y_j - Y_{j-1}|D_1 \geq j > D_0] = 0 \quad \forall j \neq j^*$.

ASSUMPTION 3* (CONCENTRATION OF COMPLIERS AT THRESHOLD): $\sum_{j \neq j^*} \Pr(D_1 \geq j > D_0) = 0$.

Assumption 2* postulates that any nonthreshold shifts in the treatment induced by the instrument have a zero average treatment effect. This is stronger than Assumption 2, which only requires this for those (off-threshold) compliers whose treatment does not cross the threshold.

Assumption 3* postulates that compliers exclusively exist at the threshold and thus experience a treatment shift from $D_0 = j^* - 1$ to $D_1 = j^*$. This is considerably stronger than Assumption 3, which allows for treatment shifts at different margins of the treatment, as long as any shift is threshold-crossing.

To see the implications of Assumptions 2* or 3*, rewrite the numerator of (3.5) as

$$\begin{aligned} E(Y|Z=1) - E(Y|Z=0) = &\sum_{j=1}^{j^*-1} E[Y_j - Y_{j-1}|D_1 \geq j > D_0] \cdot \Pr(D_1 \geq j > D_0) \\ &+ E[Y_{j^*} - Y_{j^*-1}|D_1 \geq j^* > D_0] \cdot \Pr(D_1 \geq j^* > D_0) \\ &+ \sum_{j=j^*+1}^{J} E[Y_j - Y_{j-1}|D_1 \geq j > D_0] \cdot \Pr(D_1 \geq j > D_0). \end{aligned} \tag{3.11}$$

When either Assumption 2* or 3* holds in addition to Assumption 1, (3.11) reduces to $E[Y_{j^*} - Y_{j^*-1}|D_1 = j^*, D_0 = j^* - 1] \cdot \Pr(D_1 = j^*, D_0 = j^* - 1)$ and coincides with (3.9). Furthermore, (3.8) and the denominator of (3.2) both correspond to the share of compliers at the threshold, i.e., $E(D|Z=1) - E(D|Z=0) = E(D^*|Z=1) - E(D^*|Z=0) = \Pr(D_1 = j^*, D_0 = j^* - 1)$. Therefore, $\mathrm{W}^{D^*} = \Delta^M = \Delta^* = \Delta^w = \mathrm{W}^D$. These stronger assumptions therefore allow identifying Δ^w even when the multivalued treatment is not observed.

The previous results show that when working with an instrumental variable and a binarised treatment, researchers face a classical trade-off between parameter interpretability and the strength of the identifying assumptions. While Δ^M may appear to be an attractive causal parameter due to its clear interpretation as the unit-change treatment effect for threshold compliers, it requires stronger assumptions than the weighted LATE parameter Δ^w of Angrist and Imbens (1995). The latter is identified under Assumption 1 only, but includes effects of any treatment shifts induced by the instrument. Our target parameter Δ^* lies between these two extremes in terms of identifying assumptions and complier populations. It does include effects of off-threshold shifts in the treatment, but only for compliers induced to cross the threshold.

To reiterate, Δ^* is identified under standard LATE assumptions regarding the binarised treatment only, in contrast to Δ^M. However, a clear drawback of Δ^* in terms of identification is that it includes the causal effects of off-threshold shifts in the treatment induced by the instrument, not only effects at the threshold. In the context of the returns to college, a large Δ^* might either be caused by a large threshold effect of obtaining versus just not obtaining a college degree or by off-threshold effects at different margins like attending but dropping out of college versus a high school degree. Alternatively, Δ^* might be a weighted average of both threshold and off-threshold effects. In contrast to Δ^w, however, Δ^* only includes compliers induced to cross the threshold by the instrument. Furthermore, the size of the complier groups crossing different levels of education

is identifiable, shedding some light on the size of the complier groups that enter the identified parameter Δ^*.

From the perspective of a policymaker aiming at evaluating a binarised treatment like a college degree, Δ^M seems to be an attractive parameter, corresponding to the effect of obtaining versus just not obtaining a college degree. However, if the assumptions required for the identification of this parameter appear too strong in practice, one is left with the choice between the reduced form effect of the instrument on the outcome or the treatment effects Δ^w and Δ^* for policy advice. The reduced form effect may not appear relevant in many contexts unless the instrument itself is an interesting policy. In our empirical example outlined below the reduced form is the effect of living close to a four-year college on wages and geographical distance to college is not a typical target of policy interventions. The treatment effects of education among compliers therefore appear more relevant. In contrast to Δ^w, the parameter Δ^* includes only effects of educational shifts for people who are induced to cross the policy threshold of interest, e.g., obtaining a college degree, however, at the cost of stronger identifying assumptions.

Furthermore, if interest lies in the treatment effect associated with a policy that does not exclusively affect compliers at the threshold alone, Δ^* may appear more relevant than Δ^M, precisely because it includes such policy-induced off-threshold shifts in the treatment. For instance, waiving tuition fees for students that successfully accomplish a study programme and obtain a college degree might not only induce compliers in the final year (i.e., at the threshold) to finish college but also those who would otherwise have dropped out already in earlier years (i.e., off threshold). Indeed, it is likely that many policies share the property that they shift the treatment of compliers from various treatment levels below the threshold to various levels above.

We subsequently discuss another special case of Assumption 3, which allows identifying both Δ^* and Δ^w based on D^*, even when the multivalued treatment is unobserved. As postulated in Assumption 4, we 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, which rules out compliers with other treatment margins affected. This implies that the complier population remains constant across values of j.

ASSUMPTION 4 (CONCENTRATION OF COMPLIERS AT EXTREME TREATMENT VALUES): $I\{D_1 \geq j > D_0\} = I\{D_1 \geq j^* > D_0\}$ *for all* $j, j^* \in \{1, ..., J\}$.

This assumption is stated in terms of indicator functions rather than compliance probabilities as in Assumption 3*. 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 in response to the instrument. 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. Proposition 2 states the identification of Δ^*.

PROPOSITION 2 (IDENTIFICATION OF Δ^* UNDER ASSUMPTION 4): *Under Assumptions 1 and 4, (3.11) simplifies to*

$$E(Y|Z=1) - E(Y|Z=0) = \left\{ \sum_{j=1}^{J} E[(Y_j - Y_{j-1})|D_1 \geq j^* > D_0] \right\} \cdot \Pr(D_1 \geq j^* > D_0), \tag{3.12}$$

such that $W^{D^*} = \Delta^*$.

We also note that Δ^* now 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 4, the weights in (3.3) become $\frac{\Pr(D_1 \geq j^* > D_0)}{J \cdot \Pr(D_1 \geq j^* > D_0)} = 1/J$, while in (3.2), $E(Y_j - Y_{j-1}|D_1 \geq j > D_0) = E(Y_j - Y_{j-1}|D_1 \geq j^* > D_0)$. Therefore,

$$\begin{aligned}\Delta^w &= \frac{E(Y|Z=1) - E(Y|Z=0)}{E(D|Z=1) - E(D|Z=0)} \\ &= \frac{E(Y|Z=1) - E(Y|Z=0)}{E(D^*|Z=1) - E(D^*|Z=0)} \Big/ J = \frac{\Delta^*}{J}. \end{aligned} \tag{3.13}$$

Assumption 4 thus allows identification of Δ^w using the binarised treatment, even when the underlying multivalued treatment is unobserved.

4. TESTING ASSUMPTIONS 3, 3*, AND 4

As discussed in Section 3, identification of causal effects with binarised treatments may rely on assumptions about the off-threshold treatment effects such as Assumption 2 or 2*, for which to the best of our knowledge no tests exist. In this section we propose tests for Assumptions 3, 3*, and 4, which rely on Assumption 1. We point out that if Assumption 1 is violated, this may also contribute to a rejection of the null hypotheses described below, even though the methods are not tailored to testing this assumption. We also emphasise that we test necessary, albeit not sufficient conditions for the validity of Assumptions 3 and 4. For this reason, a failure to reject the null is no proof for the satisfaction of these assumptions, even when Assumption 1 holds. A rejection, however, points to the invalidity of the respective Assumptions 3 or 4 (and/or Assumption 1, if its validity cannot be presumed), casting doubts on the IV approach using the binarised treatment unless one intends to rely on Assumption 2. In the case of Assumption 3*, however, we test both necessary and sufficient conditions, implying that, asymptotically, a nonrejection implies the satisfaction of this assumption conditional on Assumption 1.

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 maximised at the threshold. Formally, the following moment inequality constraints need to hold:

$$\begin{aligned}&\Pr(D_1 \geq j' > D_0) \geq \Pr(D_1 \geq j'' > D_0) \text{ for all } j^* \geq j' > j'' > 0, \\ &\Pr(D_1 \geq j' > D_0) \leq \Pr(D_1 \geq j'' > D_0) \text{ for all } J \geq j' > j'' \geq j^*. \end{aligned} \tag{4.1}$$

Proof. Consider the first line of (4.1) and note that

$$\begin{aligned}\Pr(D_1 \geq j' > D_0) &= \Pr(D_1 \geq j' > j'' > D_0) + \Pr(D_1 \geq j' > D_0 \geq j'') \\ &= \Pr(D_1 \geq j'' > D_0) + \Pr(D_1 \geq j' > D_0 \geq j''). \end{aligned} \tag{4.2}$$

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 (4.1) is analogous and is therefore omitted. □

By Assumption 1(a) and (b), (4.1) implies (in analogy to the discussion in (3.8) for $\Pr(D_1 \geq j^* > D_0)$) that

$$\begin{aligned} &\beta_j \geq \beta_{j'} \text{ for all } j^* \geq j > j' > 0, \\ &\beta_j \leq \beta_{j'} \text{ for all } J \geq j > j' \geq j^*, \end{aligned} \tag{4.3}$$

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 motivates the null hypothesis in Proposition 3 for testing Assumption 3 conditional on the satisfaction of Assumption 1.

PROPOSITION 3 (NULL HYPOTHESIS UNDER ASSUMPTION 3) : *Under Assumptions 1 and 3,*

$$H_0: \quad \begin{aligned} &\beta_{j+1} - \beta_j \geq 0, \text{ for all j}^* > \text{j} > 0, \\ &\beta_j - \beta_{j+1} \geq 0, \text{ for all J} > \text{j} \geq \text{j}^*. \end{aligned} \tag{4.4}$$

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 comply 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 5. To this end we reconsider the first line of (4.4) and note that

$$\begin{aligned} \beta_{j+1} - \beta_j &= \Pr(D \geq j+1 \mid Z = 1) - \Pr(D \geq j+1 \mid Z = 0) \\ &\quad - \Pr(D \geq j \mid Z = 1) + \Pr(D \geq j \mid Z = 0) \\ &= \Pr(D = j \mid Z = 0) - \Pr(D = j \mid Z = 1). \end{aligned} \tag{4.5}$$

A symmetric argument follows for the second line. Therefore, the sample analogue of (4.4) can be rewritten in the following way based on inverse probability weighting by $E(Z)$ and $1 - E(Z)$:

$$E(m_j(D, Z)) \geq 0 \tag{4.6}$$

$$\text{where } m_j(D, Z) = I\{D = j\}\frac{E(Z) - Z}{(1 - E(Z))E(Z)} \quad \text{for } j^* > j \geq 0$$

$$\text{and } m_j(D, Z) = I\{D = j\}\frac{Z - E(Z)}{(1 - E(Z))E(Z)} \quad \text{for } J > j \geq j^*.$$

These constraints match the structure of the 'cmi_test' command of Andrews et al. (2017), which verifies the sample analogue of (4.6). 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.

A rejection of (4.4) indicates the presence of nonthreshold compliance, i.e., of individuals who respond to the instrument, but are not induced to cross threshold j^*. In this case, point identification is generally lost, unless one imposes Assumption 2 or a linear IV model (using the original treatment variable D). However, researchers could still follow the path of partial identification (Manski and Pepper, 2000) and estimate upper and lower bounds on the LATE under invalid instruments (see Flores and Flores-Lagunes, 2013) or consider sensitivity checks for the robustness of the LATE under violations of the exclusion restriction (see Conley et al., 2012; Huber, 2014; Van Kippersluis and Rietveld, 2018).

Concerning Assumption 3*, both a necessary and sufficient condition for its satisfaction is that (conditional on Assumption 1) 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. This is formally stated in Proposition 4.

PROPOSITION 4 (NULL HYPOTHESIS UNDER ASSUMPTION 3*): *Under Assumptions 1 and 3*,*

$$H_0 : \beta_j = 0 \text{ for all j} \neq \text{j}^*. \tag{4.7}$$

Finally, a necessary condition for Assumption 4 is that the first stages or complier probabilities are constant across j. As highlighted in the discussion of Assumption 4 in Section 2, this implies a concentration of compliers at extreme treatment values, but is not sufficient for ruling out other complier groups. The hypothesis to be tested is given in Proposition 5.

PROPOSITION 5 (NULL HYPOTHESIS UNDER ASSUMPTION 4): *Under Assumptions 1 and 4,*

$$H_0 : \beta_j = \beta_{j+1} \text{ for all j} < \text{J}. \tag{4.8}$$

Both (4.7) and (4.8) 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, 3*, or 4. 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 (4.6), for instance, control variables can be considered by replacing $E(Z)$ everywhere with the conditional expectation of Z given the controls, also known as the 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 (4.6) within 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 it 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 (4.4) if Assumption 3 is satisfied. This in turn means that the nonsatisfaction of (4.4) 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 (4.4), (4.7), and (4.8) must hold when defining $\beta_j = \Pr(D \geq j|Z = z') - \Pr(D \geq j|Z = z'')$ for any values

Table 1. Summary statistics.

Variable	*N*	Mean	SD	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

Source: National Longitudinal Study of Young Men, 1966 and 1976 waves.

$z' > z''$ in the support of Z. For continuous instruments, the conditions given in (4.4), (4.7), and (4.8) 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, 3*, and 4 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 another variable. For instance, a test based on (4.4) may be used to verify whether a randomised labour market programme shifts labour supply only at the extensive margin (working versus not working) or also at the intensive margin (working more versus less hours). A test based on (4.7) could be applied to investigate whether participants are exclusively shifted from no to very low levels of labour supply and one based on (4.8) to test whether participants are exclusively shifted from no to full time work.

5. EMPIRICAL APPLICATION

We apply our tests to labour market data previously analysed by Card (1995) that come from the 1966 and 1976 waves of the US 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) and Mourifié and Wan (2017), we binarise the treatment to indicate having at least 16 years of education, which roughly corresponds to a four-year college degree.

The variables used in our analysis are summarised 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 binarised 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 four-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.

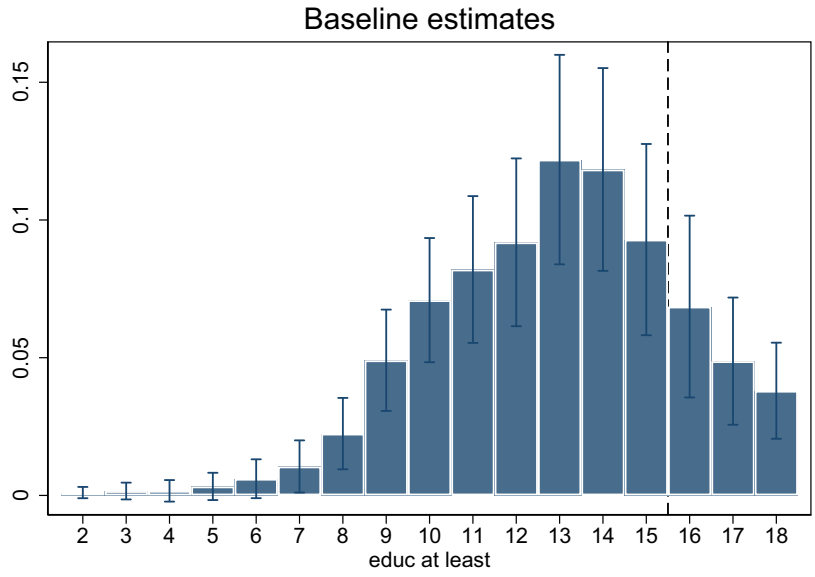

Figure 1. Effects of living close to a four-year college on years of education. Figure shows the estimated impact on binary measures of years of education equal to or above j of living close to a four-year college. The threshold for the binarised treatment is 16 or more years of education as indicated by a dashed line, corresponding roughly to a four-year college degree.
Source: Data from NLSYM.

To illustrate our tests, we first estimate the β_j parameters outlined in Section 4, 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 labour market conditions or family background, which would violate Assumption 1. As testing Assumptions 3, 3*, and 4 is conditional on Assumption 1, we control, similarly to Card (1995), for regional variables (standard metropolitan statistical area [SMSA] and region in the US) 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 a case with no controls. We observe that the pattern of coefficients are not consistent with Assumption 3*, which requires all coefficients except β_{16} to be 0. Neither does it appear to support Assumption 4, which requires the coefficients to be constant across j. Concerning Assumption 3, notice that the dashed line indicating the cut-off value for defining the binarised treatment is to the right of (rather than at) the mode of the β_j estimates, pointing to violations of the conditions in (4.1).

To formally investigate Assumption 3, we test the constraints in (4.6) using the 'cmi_test' command of Andrews et al. (2017) based on Cramer–von Mises and Kolmogorov–Smirnov test

Table 2. Tests of instrument validity with a binarised treatment.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel 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
Panel B: F-test of Assumption 3*									
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
Panel C: F-test of Assumption 4									
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

Notes: Panel A shows test statistics, critical values, and resulting p-values from tests of the moment inequalities in (4.4), 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 3*. Panel C shows F-tests of whether all β_j are the same (within cells of X), testing the special case in Assumption 4. Controls as indicated in the bottom panel. Singleton groups are dropped.

statistics.[8] The results are provided in panel A of Table 2. For interpretation, it is important to keep in mind that we test necessary, but not sufficient, conditions for the assumption to hold. Therefore,

[8] A Stata program to be found on www.github.com/martin-andresen/mvttest estimates and plots the β_j coefficients (or the maximum violations of Assumption 3 across groups, if using controls), tests Assumption 3* and 4 using F-tests, constructs the moment inequalities, and tests them using 'cmi_test'.

under Assumption 1, the rejection of the null hypothesis provides evidence against the exclusion restriction unless we impose Assumption 2, i.e., no wage effect of education among off-threshold compliers, which appears unlikely to hold. However, a nonrejection does not automatically imply the satisfaction of Assumption 3 and thus, the exclusion restriction. Without including control variables, the p-value of both the Cramer–von Mises and Kolmogorov–Smirnov statistics is 0.049, pointing to a violation of the constraints in (4.4).

When including control variables, we test for violations within cells of X, estimating the β_j coefficients in each cell and testing Assumptions 3, 3*, and 4 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 provided in columns (2)–(9) of Table 2. We find that the violations displayed in Figure 2 are statistically significant in many of the specifications, indicating the presence of off-threshold compliers in several cells of X and off-threshold levels of j. In particular, Assumption 3 is rejected in some subgroups of log wage, as seen in columns (7)–(9). This rejection is in line with findings in Mourifié and Wan (2017) who jointly test the monotonicity assumption and exclusion restriction with regard to the very same binarised college treatment based on constraints on complier densities (see Balke and Pearl, 1997). They reject the null within a specific cell of X even when adjusting p-values for multiple hypothesis testing. Conditional on the satisfaction of Assumption 1 and, thus, monotonicity, Mourifié and Wan (2017) test for the joint violation of our Assumptions 2 and 3 (implying the violation of the exclusion restriction), while our test is for Assumption 3 alone. The fact that the two conceptually quite different testing approaches both reject the null in several cells of X suggests that IV validity is likely violated for the binarised treatment, even conditional on control variables.[9]

Concerning Assumptions 3* and 4, we test the null hypotheses in (4.7) and (4.8) using F-tests in our system of equations used to estimate the β_j parameters.[10] The results are provided in panels B and C of Table 2, respectively. Both assumptions are rejected in all specifications, suggesting that compliers are neither exclusively situated at the threshold (i.e., switching from 15 to 16 years of education in response to the instrument), nor exclusively switching from the lowest to the highest level of education. Therefore, the weighted average of treatment effects based on unit changes, Δ^w, cannot be recovered based on the binarised treatment.

Overall, our results indicate that the exclusion restriction is likely violated for the binarised education measure considered unless Assumptions 2 holds. Even though the results suggest that proximity to a four-year college indeed affects education, it may do so not exclusively through 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, violating the exclusion restriction.

[9] If only one of the two tests rejected the null, IV validity would still appear doubtful, as either method can only test for necessary, but not for sufficient conditions of the respective assumptions. For this reason the tests might not simultaneously reject the null even in large samples and if both Assumptions 2 and 3 are violated.

[10] 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 heteroscedasticity.

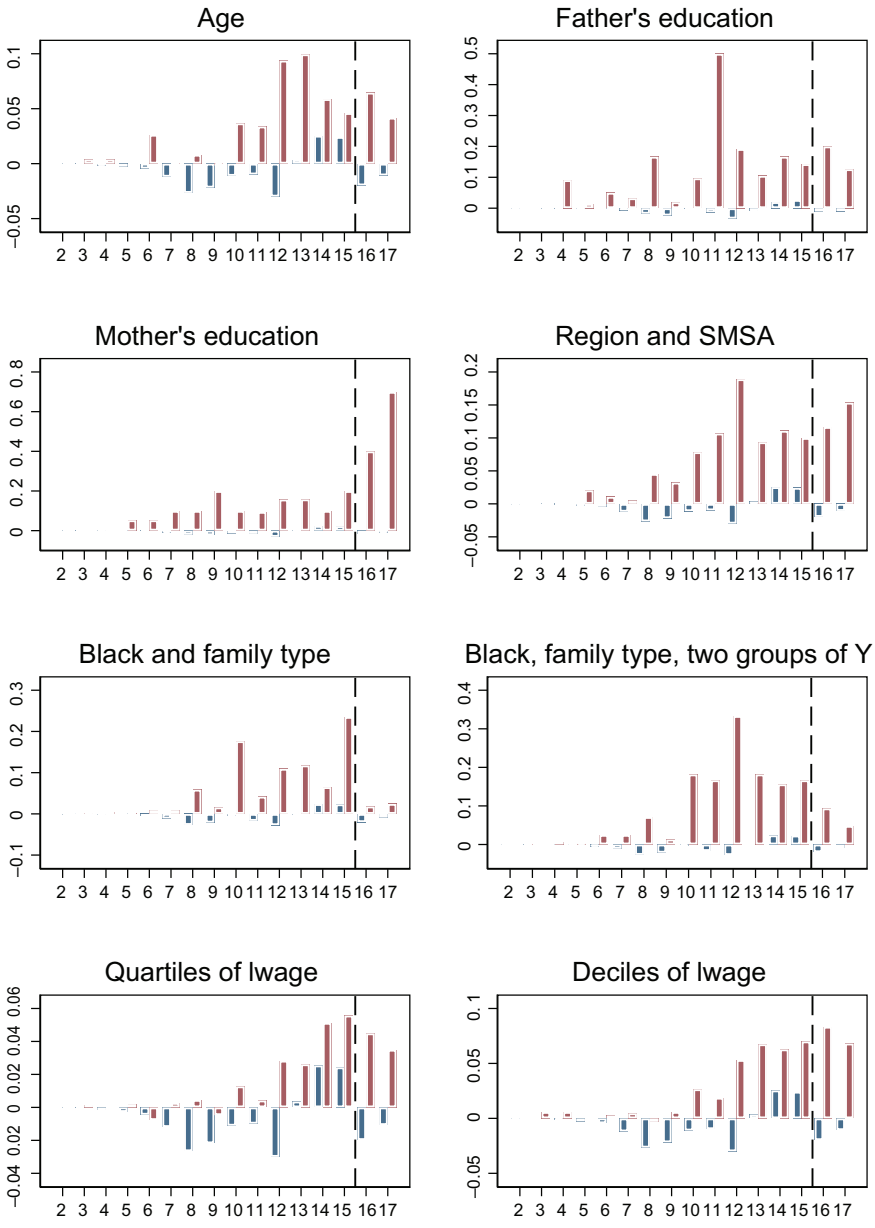

Figure 2. Maximum violations of Assumption 3 across cells. Figure shows the maximum violations across cells of X as indicated in each panel title, plotted in red Violations are $\beta_j - \beta_{j+1}$ and $\beta_{j+1} - \beta_j$ for $j \geq j^*$. For comparison, the violations in the case with no controls are plotted in blue. The threshold for the binarised treatment (j^*) is 16 or more years of education as indicated by a dashed line, corresponding roughly to a four-year college degree.

Source: Data from NLSYM.

6. CONCLUSION

In the context of IV-based estimation, we discussed threats to the exclusion restriction when binarising a multivalued endogenous treatment. Such a violation occurs whenever (a) the IV affects the multivalued treatment within support areas below and/or above the threshold for binarisation and (b) such IV-induced changes in the multivalued treatment affect the outcome. As a consequence, IV with a binarised treatment identifies the causal effect among individuals whose binary treatment complies with the IV only if either (a) or (b) 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 (a) has implications that can be tested in a moment inequality framework when the original treatment variable prior to binarisation is observed. Furthermore, when ruling out (a) and restricting the support of the multivalued treatment in a particular way, not only the average complier effect of the binarised treatment, but also a weighted average effect of unit changes of the multivalued treatment is recovered based on the binarised treatment. 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 binarised treatment generated from the multivalued years of education. Our results suggested that the exclusion restriction is violated for such a coarse definition of treatment unless Assumption 2 holds.

As a final word of caution, we emphasise that the threats to both the exclusion restriction and interpretation of identified parameters not only arise when binarising a treatment. The issues discussed in this paper prevail whenever the IV affects a finer measure of treatment than used by the researcher in their IV analysis, even when finer treatment measures are not available in the data. 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, and what parameters are identified in such a case.

ACKNOWLEDGEMENTS

We have benefited from comments from John Marshall, two anonymous referees, 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.

REFERENCES

Aizer, A. and J. J. Doyle (2015). Juvenile incarceration, human capital, and future crime: Evidence from randomly assigned judges. *Quarterly Journal of Economics 130*(2), 759–803.

Andrews, D. and X. Shi (2013). Inference based on conditional moment inequalities. *Econometrica 81*(2), 609–66.

Andrews, D. W. and X. Shi (2014). Nonparametric inference based on conditional moment inequalities. *Journal of Econometrics 179*(1), 31–45.

Andrews, D. W. K., W. Kim and X. Shi (2017). Commands for testing conditional moment inequalities and equalities. *Stata Journal 17*(1), 56–72.

Angrist, J. D. and W. N. Evans (1998). Children and their parents' labor supply: Evidence from exogenous variation in family size. *American Economic Review 88*(3), 450–77.

Angrist, J. D. 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*, 431–42.

Angrist, J. D., G. W. Imbens and D. B. Rubin (1996). Identification of causal effects using instrumental variables. *Journal of American Statistical Association 91*, 444–72(with discussion).

Balke, A. and J. Pearl (1997). Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association 92*, 1171–76.

Bhuller, M., G. B. Dahl, K. V. Løken and M. Mogstad (2020). Incarceration, recidivism, and employment. *Journal of Political Economy 128*(4), 1269–324.

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. *Quarterly Journal of Economics 120*(2), 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*, 947–52.

Card, D. (1995). Using geographic variation in college proximity to estimate the return to schooling. In Christofides, E. Grant and R. Swidinsky (Eds.), *Aspects of Labor Market Behaviour: Essays in Honour of John Vanderkamp*, 201–22, Toronto: University of Toronto Press.

Carneiro, P., J. J. Heckman and E. J. Vytlacil (2011). Estimating marginal returns to education. *American Economic Review 101*(6), 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*(1), 16–36.

Conley, T. G., C. B. Hansen and P. E. Rossi (2012). Plausibly exogenous. *Review of Economics and Statistics 94*, 260–72.

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*(6), 2356–409.

Correia, S. (2014). REGHDFE: Stata module to perform linear or instrumental-variable regression absorbing any number of high-dimensional fixed effects, Statistical Software Components S457874, Boston College Department of Economics.

Dahl, G., A. Kostøl and M. Mogstad (2014). Family welfare cultures. *Quarterly Journal of Economics 129*(4), 1711–52.

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*(2), 201–40.

Dzemski, A. and F. Sarnetzki (2014). Overidentification test in a nonparametric treatment model with unobserved heterogeneity. Conference paper, University of Mannheim, Germany.

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. Technical report, University of Sydney, Australia.

Flores, C. A. and A. Flores-Lagunes (2013). Partial identification of local average treatment effects with an invalid instrument. *Journal of Business and Economic Statistics 31*, 534–45.

Heckman, J. J. and E. Vytlacil (2001). Local instrumental variables. In Hsiao, K. Morimune and J. Powell (Eds.), *Nonlinear Statistical Inference: Essays in Honor of Takeshi Amemiya*, 1–46. Cambridge: Cambridge University Press.

Heckman, J. J. and E. Vytlacil (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*, 220–23.

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*, 467–75.

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

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*, 2043–63.

Loeffler, C. E. (2013). Does imprisonment alter the life course? Evidence on crime and employment from a natural experiment. *Criminology 51*(1), 137–66.

Manski, C. F. and J. V. Pepper (2000). Monotone instrumental variables: With an application to the returns to schooling. *Econometrica 68*(4), 997–1010.

Marshall, J. (2016). Coarsening bias: How coarse treatment measurement upwardly biases instrumental variable estimates. *Political Analysis 24*(2), 157–71.

Mogstad, M., A. Santos and A. Torgovitsky (2018). Using instrumental variables for inference about policy relevant treatment parameters. *Econometrica 86*(5), 1589–619.

Mogstad, M. and M. Wiswall (2016). Testing the quantity-quality model of fertility: Estimation using unrestricted family size models. *Quantitative Economics 7*(1), 157–92.

Mourifié, I. and Y. Wan (2017). Testing late assumptions. *Review of Economics and Statistics 99*, 305–13.

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. Working paper, University of Rochester, Rochester, NY.

Van Kippersluis, H. and C. A. Rietveld (2018). Beyond plausibly exogenous. *Econometrics Journal 21*(3), 316–31.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Replication Package

Co-editor Petra Todd handled this manuscript.

APPENDIX A: 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 A1. Defiers are ruled out by monotonicity and the population shares of never and always takers are also set to 0 (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 with regard to the total population.

Table A2 shows the results for 1,000 simulations of each of the DGPs outlined in Table A1, using 500 observations per simulation. The first column ("All") provides the results of a test of Assumption 3 based on (4.4) in the full sample. As expected, we cannot detect violations of Assumption 3 because the presence

Table A1. 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 A2. 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)
$\beta_2 - \beta_1$	−0.002	−0.003	0.0009	−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

Notes: Table reports results from 1,000 simulations of the four different data generating processes described in Table A1, using 500 observations per simulation. The reported values are the means of the respective parameters across the 1,000 simulations, standard deviations are reported in parentheses. The rejection rate is based on the 5% level of significance.

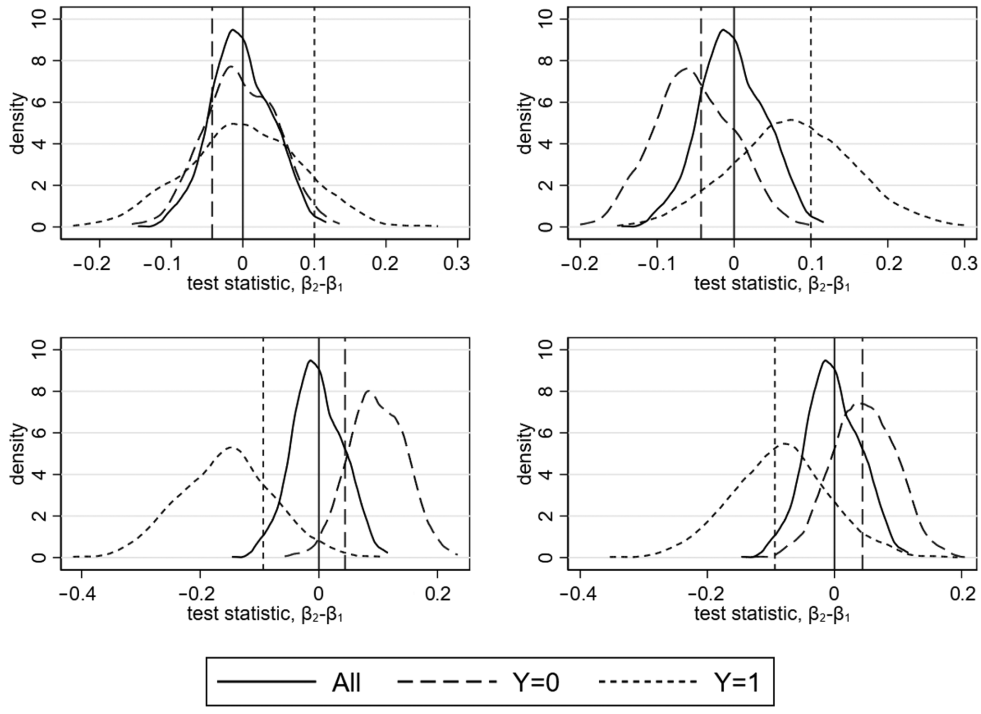

Figure A1. Density plots of test statistics. True values indicated with vertical lines.

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 nonviolating 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 A1 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.