Identifying Causal Effects of Discrete, Ordered and Continuous Treatments using Multiple Instrumental Variables

Nadja van 't Hoff*

University of Southern Denmark

November 2, 2024

Click here for the most recent version.

Abstract

Inferring causal relationships from observational data is often challenging due to endogeneity. This paper provides new identification results for causal effects of discrete, ordered and continuous treatments using multiple binary instruments. The key contribution is the identification of a new causal parameter that has a straightforward interpretation with a positive weighting scheme and is applicable in many settings due to a mild monotonicity assumption. This paper further leverages recent advances in causal machine learning for both estimation and the detection of local violations of the underlying monotonicity assumption. The methodology is applied to estimate the returns to education and assess the impact of having an additional child on female labor market outcomes.

Keywords: Ordered treatment, multiple instruments, average causal response, specification test, causal machine learning.

JEL classification: C14, C21, C26.

This project was supported by generous funding from the Independent Research Fund Denmark (90380031B). I am grateful to Giovanni Mellace for his guidance and to my fellow PhD students for their feedback. I extend my thanks to Michael Lechner, Toru Kitagawa, Volha Lazuka, Guido Imbens, Bo Honoré, Martin Huber, Phillip Heiler, Leonard Goff, and Kevin Huynh for their insightful comments. I also appreciate the feedback from the participants at "Machine Learning in Program Evaluation, High-Dimensionality, and Visualization Techniques" 2023, IAAE 2023, the seminar at Lund University in October 2023, DGPE Workshop 2023, (EC)ˆ2 2023, EWMES 2023, RES Conference 2024, the seminar at UvA in March 2024, Aarhus Workshop in Econometrics III, COMPIE 2024, the "Groningen Workshop on Causal Inference and Machine Learning" 2024, and the seminar at the University of St. Gallen in October 2024.

*Email address: navh@sam.sdu.dk.

1 Introduction

Identifying causal relationships is a central goal in economic research, but inferring causality from observational data is often challenging, particularly due to endogeneity arising from selection into treatment. Instrumental variable methods are widely used to address this issue.

Much of the existing literature focuses on identifying causal effects of binary treatments with a single instrument. Yet many real-world applications involve discrete or continuous treatments and multiple instruments can be available. For example, rather than estimating the effect of education using binary indicators like college completion, one might have data on a discrete, ordered treatment such as years of schooling. Moreover, multiple instruments might be available, such as quarter of birth (Angrist & Imbens, 1995), distance to school, or local labor market conditions (Carneiro, Heckman & Vytlacil, 2011). When treatment effects are heterogeneous, combining multiple instruments is advantageous. Each instrument generates distinct complier populations, and combining them expands the overall complier population considered, potentially bringing the local effect closer to the average treatment effect (ATE).

Two-stage least squares (TSLS) has become the standard estimation approach, but it faces two key limitations when applied to cases involving treatments with variable intensity and multiple instruments. First, the TSLS estimand is complex: Angrist and Imbens (1995) show that it represents a weighted average of average causal responses (ACRs), where each ACR reflects the effect of a one-level treatment change for specific subpopulations. In essence, the TSLS estimand consists of weighted averages of weighted averages, complicating interpretation of the effect estimates. Second, TSLS relies on a restrictive monotonicity assumption, already in the case of binary treatments. Mogstad, Torgovitsky, and Walters (2021) show that while relaxing this assumption preserves the TSLS interpretation as a weighted average, it introduces the possibility of negative weights, further complicating the effect interpretation. This issue parallels concerns raised with two-way fixed effects estimators (Borusyak & Jaravel, 2018; De Chaisemartin & d'Haultfoeuille, 2020; Goodman-Bacon, 2021), where more complex settings with relaxed assumptions result in estimands that are harder to interpret.

The main contribution of this paper is the identification of a novel causal parameter for discrete, ordered and continuous treatments with multiple instruments. This parameter features an intuitive, positive weighting scheme and is derived under a mild monotonicity assumption. Specifically, this paper introduces the *combined compliers average causal response* (CC-ACR), identified under the limited monotonicity (LiM) assumption. The CC-ACR parameter is easy to interpret, with weights that reflect the proportion of combined compliers, a sizable complier group within the population. It addresses the question, "What is the average causal effect of a one-level increase in treatment for combined compliers?", making it easier to interpret than the

TSLS estimand, while imposing a less restrictive form of monotonicity.

When treatment effects are heterogeneous, the monotonicity assumption is critical for ensuring a causally interpretable effect. In general, the monotonicity assumption restricts the direction in which the potential treatment status changes for given changes in instrument values, effectively ruling out certain response types and imposing restrictions on choice mechanisms. The CC-ACR is identified under the LiM assumption, which was originally introduced for binary treatments by van 't Hoff, Lewbel, and Mellace (2023). It imposes fewer restrictions on choice behavior than the Imbens and Angrist monotonicity (IAM) assumption (Angrist & Imbens, 1995) or the partial monotonicity (PM) assumption introduced by Mogstad, Torgovitsky, and Walters (2021). The latter was introduced in the context of binary treatments to relax the IAM assumption.

This study extends the concepts of PM and LiM to the framework of discrete, ordered or continuous treatments. Interestingly, in this context, PM shares similar limitations as IAM, which were highlighted by Mogstad, Torgovitsky, and Walters (2021) in the binary context. I show that LiM, the monotonicity assumption underlying my main identification result, not only avoids these limitations but also reduces the multiple instrument problem to a single instrument framework. In addition, it allows for the grouping of complier types, simplifying interpretation.

An additional contribution of this study is a general TSLS identification result that flexibly incorporates the various monotonicity assumptions, illustrating how each assumption influences the TSLS estimand. The challenge with TSLS lies in navigating a landscape of suboptimal options. While IAM guarantees positive weights, it is restrictive with respect to choice behavior. In contrast, PM offers greater flexibility by relaxing these restrictions to some extent, but at the risk of introducing negative weights. In both cases, the TSLS estimator converges to a weighted average of weighted averages of ACRs, and this weighting scheme complicates the interpretation of TSLS estimates.

Another key contribution of this paper is a stochastic dominance test for detecting violations of the LiM assumption, which is required for the main identification result. LiM implies positive weights for the CC-ACR parameter, which is equivalent to the condition that the cumulative distribution functions (CDFs) of the treatment, conditional on specific instrument values, do not intersect. This necessary (but not sufficient) condition can be tested empirically. Since violations of LiM in specific subgroups may average out in the full sample, a global test might fail to detect them. Building on Farbmacher, Guber, and Klaassen (2022), I show how causal forests can be used to detect such local violations by checking the sign of conditional average treatment effects within regression tree leaves, where a positive sign indicates a violation. An advantage of this method is that it enables data-driven subgroup formation in the covariate space.

To demonstrate the proposed methods, I revisit two seminal applications. The first examines

the returns to education as studied by Card (1995). Rather than focusing on the binary variable of college attendance, I focus on education measured as a discrete variable, defined by the number of years of schooling. To address potential unobserved confounding factors, such as ability, which may influence both education and wages, I follow Card (1995) and use the presence of 2-year and 4-year colleges as instruments. The combined compliers in this context are individuals whose education level increases due to living near a 2-year or 4-year college or both. The CC-ACR parameter then captures the average causal effect of an additional year of schooling for these individuals, providing a clear interpretation that cannot be obtained through TSLS. Moreover, LiM, the monotonicity assumption required for identifying the novel CC-ACR parameter, is generally more plausible than the monotonicity assumptions required for reasonable TSLS estimates. LiM accommodates individuals who prefer 2-year colleges as well as those who prefer 4-year colleges, while other assumptions do not. From a policy standpoint, understanding the impact of an additional year of schooling, rather than merely focusing on college completion, offers valuable insights for various analyses. This broader perspective can inform a range of policy decisions, including whether to extend or shorten the duration of college attendance.

Following Card (1995), I assume the instruments are exogenous, conditional on covariates such as individual and regional characteristics. I establish an identification result for the CC-ACR under conditional instrument validity, using similar arguments as Frölich (2007). A minor contribution of this paper is the extension of recent advances in causal machine learning, specifically double/debiased machine learning (DML) (Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey & Robins, 2018), from the binary treatment and single instrument setting to one with a discrete, ordered or continuous treatment and multiple binary instruments. This estimation approach has the advantage of accommodating a larger number of covariates and their interactions compared to other nonparametric estimators, which is crucial to maintain the interpretation of the CC-ACR. However, estimation is conducted on a subset of the sample, which is the cost of obtaining both a more interpretable effect estimate and a more credible monotonicity assumption.

Card (1995) primarily focuses on the 4-year college instrument, given the weakness of the 2-year college instrument. However, an advantage of the proposed methodology is that information from the weaker instrument can still be included, as weak instruments do not necessarily pose a problem when paired with a strong instrument. The results indicate that individuals attending 2-year colleges may especially benefit from additional schooling.

The second application explores the impact of an additional child on female labor market outcomes, measured as annual labor income, weekly hours worked, and weeks worked per year, following the study by Angrist and Evans (1998). Next to the classical same-sex instrument,

I use a twinning instrument that accounts for twins at any birth. In this application, LiM is more plausible than other forms of monotonicity, as it accommodates individuals who prefer same-sex as well as those who prefer mixed-sex siblings. A key advantage of the CC-ACR is that it provides insights into the average effect of having an additional child, rather than focusing solely on the effect of a third child. I find relatively small effects of an additional child on female labor market outcomes, likely due to the underlying complier population. Women with more children may have different labor preferences, which may explain the modest impact.

The remainder of this paper is organized as follows: Section 2 reviews the relevant literature. Section 3 outlines the framework, assumptions, and the CC-ACR and TSLS identification results. Section 4 details the procedure for detecting violations of the LiM assumption, while Section 5 provides guidelines for estimation. Section 6 presents the empirical findings from the two seminal studies (Card, 1995; Angrist & Evans, 1998). Finally, Section 7 offers a discussion and suggests avenues for future research. Additional results, including simulation studies for the proposed LiM test, are provided in the appendix. The code used in this study will be made publicly available on GitHub.

2 Literature review

This paper contributes to the instrumental variables literature in two key areas. First, it enhances our understanding of which causal parameters can be identified using instrumental variables. The foundation of the local average treatment effect (LATE) framework was established by Imbens and Angrist (1994) and Angrist, Imbens, and Rubin (1996). For treatments with variable intensity, Angrist and Imbens (1995) show that TSLS combines the instrument-specific weighted averages into a new weighted average. However, most literature focuses on the case of a binary treatment. In the setting with a binary treatment and multiple instruments, Mogstad, Torgovitsky, and Walters (2021) provide an identification result for TSLS under their Partial Monotonicity (PM) assumption, and van 't Hoff, Lewbel, and Mellace (2023) show that the LATE for the combined compliers is identified under their Limited Monotonicity (LiM) assumption. Goff (2024) introduces Vector Monotonicity (VM), a special form of PM, which assumes that treatment uptake is monotonic with respect to each individual instrument, rather than requiring a uniform direction of response across all instruments. He further characterizes the class of causal parameters that are point-identified under this monotonicity assumption and provides a practical two-step estimator. Frölich (2007) extends the LATE framework to include covariates nonparametrically. My findings complement those of Frölich (2007) who separately considers a discrete, ordered treatment with a single instrument, or a binary treatment with multiple instruments, while my results consider the setting with discrete, ordered treatments and

multiple instruments. My paper exhibits some connection to the work of Lee and Salanié (2018), who study discrete treatments and the point-identification of marginal treatment effects (MTE) in a framework that requires continuous instruments. Equally within the MTE framework, Heckman, Urzua, and Vytlacil (2006) consider an ordered choice model, identifying a parameter for the difference in potential outcomes between two subsequent treatment levels. Unlike the approach in the present paper, their approach requires an instrument for all incremental changes in the treatment level. Bhuller and Sigstad (2022) extend Frandsen, Lefgren, and Leslie's (2023) results for a binary treatment to a setting with multivalued treatments. They also require an instrument for every treatment level and assume no cross-effects, which is a rather restrictive assumption. Moreover, their result separately compares causal effects on specific treatment margins and does not offer an interpretation as an average effect for a one-level increase. For readers interested in a deeper exploration of instrumental variables methods that account for unobserved heterogeneity in treatment effects, I recommend the comprehensive review by Mogstad and Torgovitsky (2024).

Second, this paper contributes to the literature on specification tests of instrument validity. Research in this area has mainly been limited to joint tests on the exclusion restriction and monotonicity for a binary treatment. The first testable implications based on the exclusion and monotonicity assumptions can be traced back to Balke and Pearl (1997), Angrist and Imbens (1995), and Heckman, Urzua, and Vytlacil (2006). Angrist and Imbens (1995) show that, in case of treatments with variable intensity and a single instrument, testable implications of IAM can be established. The testable implications in the present study consider the setting with multiple instruments and the LiM assumption. There is a big strand of literature that derives results related to testable implications for joint tests in case of a binary treatment (Kitagawa, 2021; Balke & Pearl, 1997; Kitagawa, 2015; Mourifié & Wan, 2017; Huber & Mellace, 2015; Frandsen, Lefgren & Leslie, 2023; Carr & Kitagawa, 2021). Farbmacher, Guber, and Klaassen (2022) build on this literature, but employs causal forests to detect local violations of the joint assumptions, subgrouping the covariates in a data-driven way. The present paper complements this paper, as it provides a test using causal forests for LiM when the treatment is discrete and ordered. While the aforementioned literature focuses on a binary treatment, there has been some recent progress in extending the testable implications to non-binary treatment settings. For instance, Sun (2023) establishes testable implications for the exclusion and IAM assumptions for ordered and unordered nonbinary treatments.

3 Identification results

3.1 Framework and assumptions

Consider the Angrist and Imbens (1995) setup with an outcome Y, a treatment D that is discrete with bounded support, $D \in \{0, 1, ..., J\}$, such that there are $J+1$ possible treatment levels, and K binary instruments, Z_1, Z_2, ..., and Z_K. Adhering to the Rubin causal model (as detailed in Rubin, 1974, and Robins, 1986), the potential treatment states for some unit i are denoted as $D_i^{z_1 z_2 ... z_K}$, while potential outcomes are represented by $Y_i^{j, z_1 z_2 ... z_K}$.

Assumption 1: Random assignment and exclusion

$Z_k \perp\!\!\!\perp (D^{z_1 z_2 ... z_K}, Y^j) \quad \forall z_1 z_2 ... z_K \in \{0,1\}^K, k \in \{1, 2, ..., K\}, j \in \{0, 1, ..., J\}.$

Assumption 2: Stable unit treatment value assumption (SUTVA)

$Y_i^{j, z_1 z_2 ... z_K} = Y^j$ and $Y = Y^j$ if $D = j$, and

$D = D^{z_1 z_2 ... z_k}$ if $Z_1 = z_1$, $Z_2 = z_2$, ..., and $Z_K = z_K$.

Assumption 3: Instrument relevance

$0 < P(Z_1 \cdot Z_2 \cdot ... \cdot Z_K = 1) < 1$, and $0 < P((1 - Z_1) \cdot (1 - Z_2) \cdot ... \cdot (1 - Z_K) = 1) < 1$, and $P(D^{1...1...1} \geq j > D^{0...0...0}) > 0$ for some $j \in \{0, 1, ..., J\}$.

Assumption 4: Limited monotonicity (LiM)

$P(D^{1...1...1} \geq D^{0...0...0}) = 1$ or $P(D^{1...1...1} \leq D^{0...0...0}) = 1$.

The validity of the instruments relies on the independence assumption and exclusion restriction, both outlined in Assumption 1. In Card's (1995) application, the independence assumption posits that the presence of a college does not influence an individual's wage other than through the change in years of schooling attained. SUTVA (Assumption 2) ensures that the treatment level of one unit remains unaffected by the treatment level of any other unit, and that instruments assigned to a specific unit solely impact the treatment level for that particular unit. SUTVA guarantees the existence of a singular potential outcome for each treatment value. Assumption 3 requires the instruments to be relevant, which is important for estimation and for the existence of a complier population. This means that at least one instrument affects some level of the treatment to ensure the existence of compliers. For instance, this implies that the proximity to a college influences educational attainment for some individuals.

The limited monotonicity (LiM) assumption was initially introduced by van 't Hoff, Lewbel, and Mellace (2023) for the setting with a binary treatment. Assumption 4 extends LiM to settings where treatment intensity varies. It states that when exposed to all (none) of the instruments, units are at least as likely to take up treatment as when exposed to none (all) of the

instruments simultaneously. This introduces restrictions on choice behavior at the outer support of the instrument values. Without loss of generality, positive LiM ($P(D^{1...1...1} \geq D^{0...0...0}) = 1$) is assumed throughout the rest of the paper. In Card's (1995) study, this implies that an individual's educational attainment while residing close to both a 2-year college and a 4-year college is at least as large as the number of months when residing far from both a 2-year college and a 4-year college.

LiM is generally weaker than other monotonicity assumptions introduced in literature, such as the Imbens and Angrist monotonicity (IAM) assumption (Imbens & Angrist, 1994) and the partial monotonicity (PM) assumption (Mogstad, Torgovitsky & Walters, 2021). Although Vector Monotonicity (VM) as introduced by Goff (2024) is not further discussed here, as it is a special case of PM, it is important to note that it is the most empirically relevant case of PM.

IAM evaluates potential treatment states for all instrument values, basically requiring individuals to prefer one instrument over another. On the other hand, PM restricts the direction of the potential treatment status for a change in one of the instruments while keeping all other instrument values fixed. While PM has been primarily been introduced for the setting with a binary treatment, this paper extends it seamlessly to the nonbinary treatment scenario.

Imbens and Angrist monotonocity (IAM)

$P(D^{i...j...k} \geq D^{p...q...r}) = 1$ or $P(D^{i...j...k} \leq D^{p...q...r}) = 1$

$\forall\ i \in \{0,1\}, ..., j \in \{0,1\}, ..., k \in \{0,1\}$ and $\forall\ p \in \{0,1\}, ..., q \in \{0,1\}, ..., r \in \{0,1\}$

such that $P(D^{i...j...k}) \neq P(D^{p...q...r})$.

Partial monotonicity (PM)

$P(D^{1...j...k} \geq D^{0...j...k}) = 1$ or $P(D^{1...j...k} \leq D^{0...j...k}) = 1$,

$P(D^{i...1...k} \geq D^{i...0...k}) = 1$ or $P(D^{i...1...k} \leq D^{i...0...k}) = 1$, and

$P(D^{i...j...1} \geq D^{i...j...0}) = 1$ or $P(D^{i...j...1} \leq D^{i...j...0}) = 1$

$\forall\ i \in \{0,1\}, ..., j \in \{0,1\}, ..., k \in \{0,1\}$.

The restrictions on the choice mechanisms imposed by LiM are reflected in the response types of the population. In case of a binary instrument and binary treatment, Imbens and Angrist (1994) introduce the notions of always-takers, compliers, defiers, and never-takers. Here, LiM rules out the defiers. Now consider the scenario with a three-valued treatment, $D \in \{0,1,2\}$, and one binary instrument, $Z \in \{0,1\}$. There are $(J+1)^{2^K} = 3^{2^1} = 9$ initial response types. Adapting Frölich's (2007) notation, the non-responders, whose treatment level does not change in response to a change in the instrument ($D^1 = D^0$), are denoted by n_{D^1,D^0}. The compliers are the types denoted by c_{D^1,D^0} for whom $D^1 > D^0$, while for defiers, d_{D^1,D^0}, it holds that $D^1 < D^0$. Compliers are individuals for which $P(D^1 \geq D^0) = 1$, while defiers have $P(D^0 \geq D^1) = 1$. Monotonicity rules out three defier types (see Table 1).

Table 1: Initial response types with one binary instrument, $Z \in \{0,1\}$, and a three-valued treatment, $D \in \{0,1,2\}$. ✓indicates the response types allowed for under the different forms of the monotonicity assumption.

Type	D^0	D^1	LiM	PM	IAM
$c_{0,1}$	0	1	✓	✓	✓
$c_{1,2}$	1	2	✓	✓	✓
$c_{0,2}$	0	2	✓	✓	✓
$n_{2,2}$	2	2	✓	✓	✓
$n_{1,1}$	1	1	✓	✓	✓
$n_{0,0}$	0	0	✓	✓	✓
$d_{1,0}$	1	0			
$d_{2,1}$	2	1			
$d_{2,0}$	2	0			

The number of initial response types increases rapidly with the number of treatment levels. For discrete, ordered and continuous treatments, compliance intensity can vary. This means that, for a certain change in the instrument values, some response types might shift their treatment status by one level, while others might shift their treatment status by two levels. In addition, these types can have distinct baseline treatment levels, Y_i^0, adding to the complexity of types.

Next consider the scenario involving a three-valued treatment, $D \in \{0,1,2\}$, while introducing two binary instruments, $Z_1 \in \{0,1\}$ and $Z_2 \in \{0,1\}$. This amplifies the number of potential response types to $(J+1)^{2^K} = 3^{2^2} = 81$. Refer to Appendix A, Table 10, for a comprehensive listing of all initial response types. Under LiM, 54 response types remain, as types that defy with respect to the outer instrument support are eliminated, specifically types $d_{D^{00},\ldots,D^{11}}$ where $D^{00} > D^{11}$. Under PM with ordering $P(D^{10} \geq D^{00}) = 1$, $P(D^{01} \geq D^{00}) = 1$, $P(D^{01} \geq D^{11}) = 0$, $P(D^{10} \geq D^{11}) = 0$, a total of 20 response types remain. Under IAM, there are only 14 response types that remain.[1] IAM only allows for pure compliers in the sense that there are no two-way flows for any shift in the instrument values. Altogether, LiM allows for more response types and hence for rich choice heterogeneity.

In addition to allowing for rich choice heterogeneity, LiM allows us to aggregate the response types into groups, reducing the complex problem with many response types to a simpler one. Since LiM only imposes a restriction on the outer support ($Z_1 = Z_2 = 1$ and $Z_1 = Z_2 = 0$) of $\mathcal{Z} = \{(1,1),(1,0),(0,1),(0,0)\}$, the two intermediate treatment states, D^{10} and D^{01}, are

[1] For a detailed comparison of the three monotonicity assumptions when the treatment is binary see van 't Hoff, Lewbel, and Mellace (2023).

Table 2: This table presents the response types that are contained in the combined complier type $cc_{0,1}$. ✓indicates the response types allowed for under the different forms of the monotonicity assumption.

Combined type	Type	D^{00}	D^{01}	D^{10}	D^{11}	LiM	PM	IAM
$cc_{0,1}$	$c_{0,2,2,1}$	0	2	2	1	✓		
	$c_{0,1,2,1}$	0	1	2	1	✓		
	$c_{0,0,2,1}$	0	0	2	1	✓		
	$c_{0,2,1,1}$	0	2	1	1	✓		
	$c_{0,1,1,1}$	0	1	1	1	✓	✓	✓
	$c_{0,0,1,1}$	0	0	1	1	✓	✓	✓
	$c_{0,2,0,1}$	0	2	0	1	✓		
	$c_{0,1,0,1}$	0	1	0	1	✓	✓	
	$c_{0,0,0,1}$	0	0	0	1	✓	✓	

Table 3: All possible initial combined response types with two instruments, $Z_1 \in \{0,1\}$ and $Z_2 \in \{0,1\}$, and a three-valued treatment, $D \in \{0,1,2\}$. ✓indicates the response types allowed for under LiM. Combined compliers are denoted by $cc_{D^{00},D^{11}}$, combined non-responders by $cn_{D^{00},D^{11}}$, and combined defiers by $cd_{D^{00},D^{11}}$.

Combined type	D^{00}	D^{11}	LiM
$cc_{0,1}$	0	1	✓
$cc_{1,2}$	1	2	✓
$cc_{0,2}$	0	2	✓
$cn_{2,2}$	2	2	✓
$cn_{1,1}$	1	1	✓
$cn_{0,0}$	0	0	✓
$cd_{2,0}$	2	0	
$cd_{2,1}$	2	1	
$cd_{1,0}$	1	0	

not restricted. Therefore, aggregating the initial response types into response type groups is straightforward. With two instruments, I define as combined compliers, denoted as $cc_{D^{00},D^{11}}$, those types who increase the treatment level in response to changing both instrument values from zero to one ($D^{11} > D^{00}$), combined defiers, denoted as $cd_{D^{00},D^{11}}$, those types who have $D^{11} < D^{00}$, and as combined non-responders, denoted as $cn_{D^{00},D^{11}}$, those types who have $D^{11} = D^{00}$. It is important to highlight that aggregating the groups in this way is not possible under PM or IAM.

To illustrate this, consider the nine initial types in Table 2, extracted from Table 10 in Appendix A. These nine types can be aggregated into a single combined complier type, recognizing that shifting the instrument values from $(0,0)$ to $(1,1)$ increases the potential treatment status from zero to one across all nine types. This combined complier type can be denoted as $cc_{0,1}$. At the intermediate instrument values, namely $(1,0)$ and $(0,1)$, this aggregated type can respond as complier or defier with respect to either instrument. In a similar fashion, the remaining response types in Table 10 can be aggregated into groups, effectively reducing the initial 81 types to the nine aggregated types showcased in Table 3. Similar results apply to settings where the treatment attains more than three levels or where more than two binary instruments are available. LiM naturally reduces a complex setting to a simple comparison between two different potential treatment states, $D^{1\ldots1\ldots1}$ and $D^{0\ldots0\ldots0}$, independently of the number of instruments. Notably, combined defier types with $cd_{a,b}$ where $b < a$ are ruled out by the LiM assumption, but all other defier types are not.

3.2 The combined compliers ACR

Theorem 1 provides the main result, namely the CC-ACR, a novel causal parameter which has a straightforward interpretation and is derived under the LiM assumption.

Theorem 1: The combined compliers average causal response (CC-ACR)
Let Assumptions 1[2]*, 2, 3, and 4 hold. Then a weighted average of average causal responses for the combined complier subpopulations is identified:*

$$\begin{aligned}\beta_{\text{CC-ACR}} &\equiv \frac{E(Y|Z_1 = Z_2 = ... = Z_K = 1) - E(Y|Z_1 = Z_2 = ... = Z_K = 0)}{E(D|Z_1 = Z_2 = ... = Z_K = 1) - E(D|Z_1 = Z_2 = ... = Z_K = 0)} \\ &= \sum_{k<l} \frac{(l-k)\cdot P(T = cc_{k,l})}{\sum_{m<h}(h-m)\cdot P(T = cc_{m,h})} \cdot E\left(\frac{Y^l - Y^k}{l-k}|T = cc_{k,l}\right).\end{aligned} \tag{1}$$

Proof in Appendix B.1.

[2]Assumption 1 can be relaxed to hold only for the instrument $\widetilde{Z}$, where $\widetilde{Z} = 1$ if $Z_1 = Z_2 = \cdots = Z_K = 1$ and $\widetilde{Z} = 0$ if $Z_1 = Z_2 = \cdots = Z_K = 0$, but the original form of Assumption 1 is equally plausible in most cases.

T denotes type and the set of response types, cc, consists of the combined complier types denoted $cc_{k,l}$ where $l > k$. These are the complier types that increase their treatment level in response to shifting all instruments from zero to one. Theorem 1 states that a weighted average of causal responses, $E(Y^l - Y^k)$, that are scaled by the change in treatment level, $(l - k)$, over these combined complier subpopulations is identified. It should be noted that this identification result is robust to the presence of non-responders. It is further important to emphasize that the weights of the CC-ACR are always positive by construction and sum up to one.

Theorem 1 takes into account that the treatment responses vary in intensity. Within the context of the returns to education as considered by Card (1995), the CC-ACR provides a weighted average of causal responses for individuals who extend their education when all instrument values shift from zero to one. In this case, all instrument values equaling one indicates that individuals reside close to both a 2-year and a 4-year college. To clarify Theorem 1, note that $E(Y^{14} - Y^{12} \mid cc_{12,14})$ measures the average effect on an individual's wage when obtaining 14 instead of 12 years of schooling, for those who adjust their education level accordingly in response to this change in instrument values. This average effect is weighted by the probability of belonging to this complier type, represented by $P(T = cc_{12,14})$, providing weights proportional to the response type group size. Further, $E(Y^{14} - Y^{12} \mid cc_{12,14})$ represents the effect of 2 additional years of schooling, reflected in scaling the difference in outcomes, $Y^{14} - Y^{12}$, by the treatment level difference, $(l - k) = 2$ years.

Theorem 1 simplifies if treatment effects are linear, meaning the impact of increasing schooling by one year is equivalent, whether it is from 12 to 13 years or from 15 to 16 years. In this case, the CC-ACR can be interpreted as the average effect of a one-level increase among the combined complier population: $E(Y^j - Y^{j-1} \mid T \in cc)$, without considering the treatment margins involved. The following corollary demonstrates the interpretation of treatment effects within this linear framework.

Corollary 1: Linearity of treatment effects

Let Assumptions 1, 2, 3, and 4 hold. Under linearity of the treatment effects, it holds for every treatment level, $j \in 1, ..., J$, that

$$\begin{aligned}\beta_{\text{CC-ACR}} &\equiv \frac{E(Y|Z_1 = Z_2 = ... = Z_K = 1) - E(Y|Z_1 = Z_2 = ... = Z_K = 0)}{E(D|Z_1 = Z_2 = ... = Z_K = 1) - E(D|Z_1 = Z_2 = ... = Z_K = 0)} \\ &= \sum_{k<l} \frac{(l-k) \cdot P(T = cc_{k,l})}{\sum_{m<h}(h-m) \cdot P(T = cc_{m,h})} \cdot E\left(Y^j - Y^{j-1} | T = cc_{k,l}\right).\end{aligned} \tag{2}$$

The following illustrative example clearly shows how Corollary 1 emerges from Theorem 1:

$$E\left(\frac{Y^2-Y^0}{2-0}|T=cc_{0,2}\right)=E\left(\frac{Y^1-Y^0}{2}|T=cc_{0,2}\right)+E\left(\frac{Y^2-Y^1}{2}|T=cc_{0,2}\right)$$
$$=2\cdot E\left(\frac{Y^1-Y^0}{2}|T=cc_{0,2}\right)=E(Y^1-Y^0|T=cc_{0,2})=E(Y^2-Y^1|T=cc_{0,2}).$$

This example illustrates that under linear treatment effects, the expected difference in the outcome when changing the treatment status from zero to one is equivalent to the expected difference when changing the treatment status from one to two for the combined compliers of type $cc_{0,2}$. These are the response types that would change their treatment status from zero to two when all instruments are changed from zero to one.

3.3 Identification of the CC-ACR including covariates

The result presented in the preceding section did not consider identification in the presence of relevant covariates. However, numerous real-world applications exist where the instruments are only valid after conditioning on covariates. Taking the returns to education application as an example, one might worry about factors influencing an individual's surroundings and their wage. The presence of a college might correlate with various individual and county characteristics. Thus, Assumption 1 must be adjusted so that the instruments are approximately randomly assigned conditional on these characteristics.

Assumption 1C: Unconfoundedness and exclusion

$$Z_k\perp(D^{z_1z_2...z_K},Y^j)|X \qquad \forall z_1,z_2,...,z_K,k\in\{1,2,...,K\},j\in\{0,1,...,J\}.$$

In addition to Assumption 1C and Assumptions 2 to 4, common support is assumed to guarantee that there is overlap in the observed characteristics at the outer support of the instrument distribution.

Assumption 5: Common support

$$P(X=x|Z_1=z_1,Z_2=z_2,...,Z_K=z_k)>0 \qquad \forall x\in\mathcal{X},\forall z_1z_2...z_K\in\{0,1\}^K.$$

Theorem 1 can easily be extended to hold conditional on covariates:

$$\begin{aligned}\beta_{\text{CC-ACR}}(X)&=\frac{E(Y|X,Z_1=Z_2=...=Z_K=1)-E(Y|X,Z_1=Z_2=...=Z_K=0)}{E(D|X,Z_1=Z_2=...=Z_K=1)-E(D|X,Z_1=Z_2=...=Z_K=0)}\\&=\sum_{k<l}\frac{(l-k)\cdot P(T=cc_{k,l}|X)}{\sum_{m<h}(h-m)\cdot P(T=cc_{m,h})}E\left(\frac{Y^l-Y^k}{l-k}|X,T=cc_{k,l}\right).\end{aligned} \tag{3}$$

Then, to obtain $\beta_{\text{CC-ACR}}$, integrating over the distribution function $f_{x|\text{combined complier}}(x)$ is required. This function is unknown, but following Frölich (2007) and using Bayes' theorem, it

is straightforward to show that $f_{x|\text{combined complier}}(x)$ equals the estimable distribution function f_x, weighted with the corresponding increments of in the treatment level, $(l-k)$:

$$f_{x|\text{combined complier}}(x) = \frac{\sum_k^K \sum_{k<l}^K P(T = cc_{k,l}|X) \cdot (l-k)}{\sum_k^K \sum_{k<l}^K P(T = cc_{k,l}) \cdot (l-k)} \cdot f_x(x).$$

This allows for identification of the CC-ACR under Assumption 1C, as formalized in Corollary 2.

Corollary 2: The CC-ACR under unconfoundedness

Let Assumptions 1C and 2 to 5 hold. Then, the CC-ACR is given by

$$\begin{aligned}
&\beta_{\text{CC-ACR}} \\
&= \int \beta(x) \cdot f_{x|\text{combined complier}}(x)dx \\
&= \frac{\int (E(Y|X=x, Z_1 = Z_2 = ... = Z_K = 1) - E(Y|X=x, Z_1 = Z_2 = ... = Z_K = 0)) \cdot f_x(x)dx}{\int (E(D|X=x, Z_1 = Z_2 = ... = Z_K = 1) - E(D|X=x, Z_1 = Z_2 = ... = Z_K = 0)) \cdot f_x(x)dx}.
\end{aligned}$$

For brevity, define

$$\widetilde{Z} = \begin{cases} 1 & \text{if } Z_1 = Z_2 = ... = Z_K = 1 \\ 0 & \text{if } Z_1 = Z_2 = ... = Z_K = 0 \end{cases}.$$

Considering only the subsample at the outer support of the instrument distribution with $\widetilde{Z}$ as the sole instrument, this expression reduces to

$$\beta_{\text{CC-ACR}} = \frac{\int (E(Y|X=x, \widetilde{Z}=1) - E(Y|X=x, \widetilde{Z}=0)) \cdot f_x(x)dx}{\int (E(D|X=x, \widetilde{Z}=1) - E(D|X=x, \widetilde{Z}=0)) \cdot f_x(x)dx}. \tag{4}$$

3.4 The causal interpretation of two-stage least squares

In this section, I extend prior research by Mogstad, Torgovitsky, and Walters (2021), which primarily delves into the causal interpretation of two-stage least squares (TSLS) with a focus on a binary treatment and multiple, mutually-exclusive instruments under the PM assumption. The goal is to generalize this result to the broader context of a discrete, ordered or continuous treatment, without imposing monotonicity at first. The probability limit of TSLS is given by Proposition 1.

Proposition 1: The causal interpretation of TSLS

Let M denote the number of elements in the rectangular instrument support $\mathcal{Z} = \{z_0, ..., z_l, ..., z_m\}$, ordered such that $l < m$ implies $E(D|Z = l) < E(D|Z = m)$. Let $I(\cdot)$ denote the indicator function, which equals one if its argument is true and zero otherwise. Suppose that Assumptions 1, 2, and 3 are satisfied. Then,

$$\beta_{\text{TSLS}} = \sum_{t \in \mathcal{T}_{\text{M}}} P(T = t) \sum_{m=1}^{M} \iota_{m,m-1} \cdot \omega_m \cdot E(Y^{D^{z_m}} - Y^{D^{z_{m-1}}} | T = t), \tag{5}$$

where

$$\omega_m = \frac{(1 - P(Z \geq z_m))P(Z \geq z_m) \cdot \{E(D|Z \geq z_m) - E(D|Z < z_m)\}}{\sum_{l=0}^{M} P(Z = z_l)E(D|Z = z_l)(E(D|Z = z_l) - E(D))},$$

and

$$\iota_{m,m-1} \equiv I(D^{z_m} \geq D^{z_{m-1}}) - I(D^{z_m} \leq D^{z_{m-1}}),$$

where $\mathcal{T}_{\text{M}}$ is the set of response types that are allowed for under the specified monotonicity assumption.

Proof in Appendix B.2.

Proposition 1 reveals that TSLS gives a weighted average of average causal responses (ACR), $E(Y^{D^{z_m}} - Y^{D^{z_{m-1}}})$, corresponding to the response types, t, present in the population. The weights determine the contribution of each local average causal response to the parameter β_{TSLS}. Similar to the CC-ACR, the weights consist of $P(T = t)$, the probability of observing a certain response type, and are non-negative and sum to one. However, the TSLS estimand contains additional, rather arbitrary weighting terms. Consider, for instance, the weights ω_m. These weights are proportional to $P(Z \geq z_m)(1 - P(Z \geq z_m))$, effectively giving more weight to $E(Y^{D^{z_m}} - Y^{D^{z_{m-1}}} | T = t)$ when it lies in the center of the instrument distribution. It is hard to come up with an empirical setting where this is a desirable feature of the TSLS weights. Ordering the values of the instrument support $\mathcal{Z} = \{z_0, ..., z_l, ..., z_m\}$, such that $l < m$ implies $E(D|Z = l) < E(D|Z = m)$, results in $(E(D|Z \geq z_m) - E(D|Z < z_m))$ being positive for all z_m. Note that this implies that the constructed instrument should be monotonic with the propensity score to ensure non-negative weights ω_m. This expression shows that more weight is given if comparatively more types respond to a change in the instrument values. Next to the weight ω_m, the TSLS estimand contains the term $\iota_{m,m-1}$, which can attain three values: $\iota_{m,m-1}$ equals 1 when $D^{z_m} > D^{z_{m-1}}$, it equals -1 when $D^{z_m} < D^{z_{m-1}}$, and 0 when $D^{z_m} = D^{z_{m-1}}$. In the latter case, it holds that $E(Y^{D^{z_m}} - Y^{D^{z_{m-1}}} | T = t) = 0$. $\iota_{m,m-1}$ guarantees the interpretation of a weighted average of causal responses $Y^a - Y^b$ for which $a > b$. Simply put, it switches $E(Y^{D^{z_m}} - Y^{D^{z_{m-1}}})$ to $E(Y^{D^{z_m - 1}} - Y^{D^{z_m}})$ whenever $D^{z_m - 1} > D^{z_m}$.

Proposition 1 is derived without imposing any form of the monotonicity assumption. The advantage is that it enables researchers to reflect upon the response types existing in the

population, drawing on prior knowledge or subject expertise. Proposition 1, therefore, provides flexibility in formulating monotonicity. Without imposing monotonicity, β_{TSLS} contains negative weights because $\iota_{m,m-1}$ is always less than or equal to 1 for some response types. This results in a negative weight on the local ACRs for these types, reversing the sign of these specific ACRs in the weighted average of ACRs. Imposing certain forms of monotonicity can eliminate those types for which sign reversals in the average local effects occur. For instance, IAM only allows for types which never have $\iota_{m,m-1} = -1$, thus guaranteeing positive weights. However, IAM does have the drawback of ruling out numerous response types, thereby heavily constraining choice heterogeneity. PM is often more reasonable to assume, given that it allows for a broader range of response types. Under PM, response types are allowed to be present even when $\iota_{m,m-1} = -1$ for some m, provided that they have $\iota_{m',m'-1} = 1$ for some other m'. That is, PM ensures that the allowed response types also respond with an increase in the treatment level for some change in instrument values such that the local causal responses in the expression $\sum_{m=1}^{M} \iota_{m,m-1} \cdot \omega_m \cdot E(Y^{D^{z_m}} - Y^{D^{z_{m-1}}} | T = t)$ for this type t are not all negatively weighted. It is alarming that, if the defier responses outweigh the complier responses, the ACR for this type is negatively weighted.

In short, researchers face a choice between imposing the more restrictive IAM assumption, which guarantees positive weights, or adopting PM with the risk of introducing sign-reversal issues through the weights $\iota_{m,m-1}$. In either scenario, the weights ω_m remain arbitrary and lack intuition. Moreover, PM might still be too restrictive in certain applications. Notably, the CC-ACR parameter outlined in Theorem 1 bypasses these concerns: its weights are positive by construction and it is identified under the generally weaker LiM assumption. Note that under LiM, the TSLS estimand will always include some negative weights.

3.5 Alternative representation of the CC-ACR

Theorem 1 above introduced the CC-ACR, expressed as a weighted average of causal responses for different response types. This representation is extremely valuable for understanding the influence of the response types on the interpretation of this causal parameter. This section offers an alternative representation of the CC-ACR.[3] The representation presented here shifts the focus to changes in the treatment status and is expressed in terms of an average of responses along the causal response function. The causal response function is given by the sequence $Y^j - Y^{j-1}$ for every unit.[4] While the representation of the CC-ACR in this section is equivalent

[3] In a similar fashion, the TSLS estimand has an alternative representation, as shown in Appendix B.4.

[4] As mentioned by Angrist and Imbens (1995), with $J + 1$ treatment levels, the $\frac{(J+1)J}{2}$ potential treatment effects can be written with respect to the J linearly independent treatment effects when the treatment level increases with one unit: $Y^j - Y^{j-1}$. Thus, if $D \in \{0, 1, 2\}$, then $J + 1 = 3$, meaning that there are six possible

in terms of its definition, it offers a different perspective on the interpretation.[5] Of particular importance, the representation of the CC-ACR in this section introduces a weighting function that offers two advantages over the weighting function of the CC-ACR of Theorem 1: First, the weights can be estimated, and second, they also serve as a means for conducting validity tests on the LiM assumption, as will be demonstrated later in Section 4.

Proposition 2: Alternative representation of Theorem 1

Suppose that the same conditions hold as in Theorem 1. Then,

$$\begin{aligned}\beta_{\text{CC-ACR}} &\equiv \frac{E(Y|Z_1 = Z_2 = ... = Z_K = 1) - E(Y|Z_1 = Z_2 = ... = Z_K = 0)}{E(D|Z_1 = Z_2 = ... = Z_K = 1) - E(D|Z_1 = Z_2 = ... = Z_K = 0)} \\ &= \sum_{j=1}^{J} \frac{P(D^{1...1...1} \geq j > D^{0...0...0})}{\sum_{i=1}^{J} P(D^{1...1...1} \geq i > D^{0...0...0})} \cdot E(Y^j - Y^{j-1}|D^{1...1...1} \geq j > D^{0...0...0}).\end{aligned} \tag{6}$$

Proof in Appendix B.3.[6]

The weights lie between zero and one, and collectively sum up to one. The weights depend on the relative strength of the instruments, which is closely tied to the corresponding proportions of combined compliers. Compliers whose treatment level is moved along multiple treatment levels by the instrument contribute multiple times in the weights, which is equivalent to the representation in Theorem 1. The connection between the two representations can be seen as follows: $P(D^{1...1...1} \geq j > D^{0...0...0}) = \sum_{k<j\leq l} P(T = cc_{k,l})$.

To illustrate the general result of the alternative representation of Theorem 1, consider the example of a three-valued treatment, $D \in \{0, 1, 2\}$, and two binary instruments, $Z_1 \in \{0, 1\}$ and $Z_2 \in \{0, 1\}$. Note that the combined compliers denoted by $cc_{2,0}$ increase their treatment level from zero to two when all instrument values change from zero to one, and hence these compliers contribute twice to the alternative representation in Equation (6). This can be seen from the weights: These particular compliers contribute to both $P(D_i^{11} \geq 2 > D_i^{00})$ and $P(D_i^{11} \geq 1 > D_i^{00})$. Conversely, compliers that respond with a one-level change in the treatment due to this change in instrument values contribute only once. The compliers in the aggregated complier group $cc_{2,1}$ contribute to $P(D_i^{11} \geq 2 > D_i^{00})$, while the compliers in the aggregated complier group $cc_{1,0}$ contribute to $P(D_i^{11} \geq 1 > D_i^{00})$. The usefulness of the weighting function in Equation (6) is detailed further in the next section.

treatment effects: $Y^1 - Y^0$, $Y^2 - Y^0$, $Y^2 - Y^1$, $Y^0 - Y^1$, $Y^0 - Y^2$, and $Y^1 - Y^2$.

[5]The representation in Theorem 1 resembles the result by Frölich (2007), whereas the representation in this section more closely resembles the result by Angrist and Imbens (1995).

[6]The proof relies on D consisting of integer values in $\{0, 1, ..., J\}$, and can in some settings be obtained by a linear transformation of D, which boils down to multiplying the CC-ACR with a constant.

3.6 Weighting function of the alternative representation

The alternative formulation of Theorem 1, as presented in Equation (6) in the previous section, offers two substantial advantages. The first advantage is that it allows for a better understanding of the estimates as it permits the estimation of the weights in Equation (6), since

$$\begin{aligned}
&P(D^{1...1...1} \geq j > D^{0...0...0}) \\
&= P(D^{1...1...1} \geq j) - P(D^{0...0...0} \geq j) \\
&= P(D^{0...0...0} < j) - P(D^{1...1...1} < j) \\
&= P(D < j | Z_1 = Z_2 = ... = Z_K = 0) - P(D < j | Z_1 = Z_2 = ... = Z_K = 1)^7,
\end{aligned} \tag{7}$$

where $P(D^{0...0...0} < j) - P(D^{1...1...1} < j) = P(D < j | Z_1 = Z_2 = ... = Z_K = 0) - P(D < j | Z_1 = Z_2 = ... = Z_K = 1)$ holds because of independence.

It is important to note that the weights given by the group type shares $P(T = cc_{k,l})$ of the CC-ACR estimand in Equation (1) are not point-identified, meaning that these weights cannot be estimated without imposing additional assumptions. Consider, for example, the simplest case where $P(D^{11} \geq 1 > D^{00}) = P(T = cc_{0,1}) + P(T = cc_{0,2})$ and $P(D^{11} \geq 2 > D^{00}) = P(T = cc_{1,2}) + P(T = cc_{0,2})$. These are two equations with three unknowns. Ruling out one type would allow for point-identification and estimation of the shares.

The second advantage of deriving the weights in Equation (6) is that necessary conditions for the validity of the LiM assumption arise from the weighting function. If LiM (i.e., $P(D^{1...1...1} \geq D^{0...0...0}) = 1$) holds, then it must hold that $P(D^{1...1...1} \geq j) - P(D^{0...0...0} \geq j) \geq 0$ for all j. Consequently, the expression in Equation (7) must be greater than or equal to zero under LiM, meaning that the LiM assumption implies that the weighting function is positive across all treatment levels and vice versa.

3.7 Identification of the CC-ACR for a continuous treatment

Up to this point, the present study has focused on discrete, ordered treatments. However, it is worth noting that in many settings the treatment can be continuous. Theorem 2 provides the results for the continuous treatment setting.

[7] Angrist and Imbens (1995) provide the weights for the setting with a single binary instrument.

Theorem 2: Continuous treatment effect

Let Assumptions 1, 2, 3, and 4 hold. If the treatment is continuous, the CC-ACR is identified as

$$\beta_{\text{CC-ACR}} = \frac{E(Y|Z_1 = Z_2 = ... = Z_K = 1) - E(Y|Z_1 = Z_2 = ... = Z_K = 0)}{E(D|Z_1 = Z_2 = ... = Z_K = 1) - E(D|Z_1 = Z_2 = ... = Z_K = 0)}$$
$$= \int_0^\infty \frac{P(D^{1...1...1} \geq t > D^{0...0...0})}{\int_0^\infty P(D^{1...1...1} \geq j > D^{0...0...0})dj} \cdot \frac{\partial E(Y^t|D^{1...1...1} \geq t > D^{0...0...0})}{\partial t} dt.$$

Proof in Appendix B.5.

Theorem 2 is analogous to the CC-ACR derived for discrete, ordered treatments in Equation (6). It essentially represents a weighted average derivative, with the weighting terms being determined by the shifts among combined compliers resulting from simultaneously moving all instrument values from zero to one. The included combined complier types are those types whose treatment status t lies between $D^{1...1...1}$ and $D^{0...0...0}$. Hence, the stronger the instrument, the larger the subpopulation considered by the CC-ACR. The weight assigned to the specific potential treatment levels is proportional to the share of complier types whose treatment status t lies between $D^{1...1...1}$ and $D^{0...0...0}$.

3.8 Binarizing continuous instruments

In some applications, instruments may be continuous and can be binarized to apply the proposed methodology. The choice of cut-off value is critical, as it significantly affects the interpretation of the CC-ACR estimand by altering the complier population.

In certain cases, there is a clear rationale for selecting a cut-off, particularly in economic analyses focused on the average causal effect within a specific complier group. For example, when evaluating the impact of a new college, policymakers might set a proximity threshold, coding distances under ten minutes as zero and those over sixty minutes as one. This approach helps isolate the LATE for individuals responsive to this threshold, aiding targeted policy development.

Without such a theoretical basis, one might aim to estimate the causal effect for the largest feasible complier group. Selecting observations with extreme propensity scores captures substantial shifts in treatment assignment but reduces the number of observations. Conversely, using the median as a cut-off retains more observations but captures a smaller portion of the complier group. This trade-off requires careful consideration and a balanced approach is advisable. In the absence of economic guidance, setting the cut-off at the 25th and 75th percentiles might offer a practical compromise, likely preserving more observations while capturing a larger complier group than a median-based cut-off.

4 Test for detecting violations of LiM

In this section, I show how the necessary conditions implied by LiM, derived in Section 3.6, can be exploited to construct formal statistical tests for detecting violations of the LiM assumption.

4.1 Global violations

As demonstrated in Section 3.6, under LiM, it must hold that the CDF of D given $Z_1 = Z_2 = ... = Z_K = 1$ and the CDF of D given $Z_1 = Z_2 = ... = Z_K = 0$ do not cross, and the former CDF first-order stochastically dominates the latter CDF. This is a necessary (though not sufficient) condition that can be verified from the data.[8] Global violations can be quickly detected through visual inspection: if the CDFs do not intersect, the necessary condition for LiM holds across all instances of the causal response function. A more formal testing procedure for stochastic dominance can be obtained through the Kolmogorov-Smirnov test, or through a multiplier bootstrap test, which is particularly of interest when the asymptotic distribution of the test statistic under the null hypothesis is unknown (see, for example, Abadie (2002)). A formal global test lies outside the scope of this paper. Instead, the next section focuses on detecting local violations.

4.2 Local violations

This section addresses the possibility that more severe local violations of LiM could exist within specific subgroups and get averaged out in the full sample, which would decrease power of a test for LiM. It can be shown that LiM implies that the following inequality must be satisfied at any point x in the covariate space (see Appendix C.1):

$$E(I(D < j)|\widetilde{Z} = 0, X = x) - E(I(D < j)|\widetilde{Z} = 1, X = x) \geq 0 \text{ for all } j \in \{0, 1, ..., J\}, \quad (8)$$

where $\widetilde{Z}$ equals one if $Z_1 = Z_2 = ... = Z_K = 1$ and zero otherwise. Testing a condition at every level of the treatment, j, can offer the advantage of detecting for which causal response, $E(Y^j - Y^{j-1}|D^{1...1...1} \geq j > D^{0...0...0})$, the weight of τ_j in Equation (6) might be negative. This provides knowledge at what point of the causal response function the assumption might be violated. A disadvantage is that it can be computationally intensive, especially if the treatment can attain a large range of values.

The necessary conditions established in Expression (8) boil down to estimating heterogeneous causal effects of the instrument on the treatment. Let (D_i, Z_i, X_i) be i.i.d. observations for

[8] Note that, in a similar fashion, one can consider the choice restrictions imposed by PM and verify that these hold across the treatment margins by comparing conditional CDFs. While LiM can be tested with only one comparison, PM involves $K \cdot 2^{K-1} = 12$ comparisons of CDFs when $K = 3$ instruments are available.

$i = 1, ..., n$, and define the pseudo variable $Q_{j,i} \equiv I(D_i < j)$. Then, write the conditional average treatment effect (CATE) of $\widetilde{Z}$ on Q_j at the point $X = x$ as

$$\tau_j(x) = E(Q_{j,i}|\widetilde{Z} = 1, X_i = x) - E(Q_{j,i}|\widetilde{Z} = 0, X_i = x). \tag{9}$$

Under Assumptions 1 to 4, the inequality $\tau_j(x) \leq 0$ has to be true for every combination of j and x. If $\tau_j(x)$ is positive, this indicates a violation of LiM, meaning that the necessary conditions for LiM can be interpreted as learning the sign of a conditional average treatment effect (CATE). Moving forward, I closely follow the procedure proposed by Farbmacher, Guber, and Klaassen (2022) (see Appendix C.2 for a detailed description). First, a causal forest (Wager & Athey, 2018) is employed for estimating these heterogeneous CATEs. Then, the heterogeneity is summarized by shallow Breiman trees, which additionally allow for visualization of the test. Relevant subgroups are selected through pruning of these trees. Finally, promising subgroups with potentially positive CATEs are selected and tests with Bonferroni-corrected p-values are performed.[9]

The proposed approach offers two main benefits. Firstly, if the degree of violation of the assumption differs for different subgroups, then this test has larger power than alternative tests, since it checks for violations of monotonicity in a specific area of the covariate space instead of in the full sample. Secondly, it is beneficial to form subgroups in a data-driven way, instead of having a researcher create potentially arbitrary subgroups. The latter can be especially inefficient in case of high-dimensionality of the covariate space.

A violation of LiM can have substantial consequences, potentially leading to estimating the wrong sign of the CC-ACR parameter or to less precise estimates. These issues are particularly pronounced in the case of few compliers (Angrist, Imbens & Rubin, 1996). LiM does allow for response types that defy with respect to some of the instruments, as long as they can be pushed towards compliance by some other instrument. However, defier types that respond most strongly to the instrument that they defy are problematic. The presence of these defiers exacerbates this bias, which is influenced by the instrument's strength and the variability of treatment effects. Insights into the magnitude of the violation can be gained through sensitivity tests.[10] Detecting a violation within any subpopulation undermines the IV validity for the whole population, as

[9] In some applications, the Bonferroni correction might be too conservative and exhibit low power. This correction is most effective when tests are independent, which is clearly not the case here. Consequently, it might be more beneficial to calculate the critical value while considering the correlation between variables and tests. Chernozhukov, Chetverikov, Kato, and Koike (2023) propose a bootstrap approach that allows to test on uniformly valid confidence intervals. In an effort to increase power, Huber and Kueck (2022) implement a multiplier bootstrap which involves a score function with dimensions equal to the number of leaves tested.

[10] For instance, Klein (2010) offers insights into recovering the LATE when monotonicity violations occur randomly and how to approximate the bias. Noack (2021) develops methods to assess the sensitivity of LATEs to violations of IAM. Extending these findings to LiM violations presents an intriguing avenue for future research.

it raises the potential for similar issues to exist in other subpopulations (Farbmacher, Guber & Klaassen, 2022).

5 Estimation of the CC-ACR

5.1 Estimation without covariates

There are several ways of estimating the CC-ACR presented in Equation (1). A simple approach is to implement the TSLS method within the subsample of observations at the outer support of the instrument values using $\widetilde{Z} = Z_1 = Z_2 = ... = Z_K$ as the single instrument. Assuming independent sampling, this strategy yields consistent estimates and asymptotically valid confidence intervals for the parameter $\beta_{\text{CC-ACR}}$. An alternative estimation approach involves formulating moment equations and subsequently utilizing the generalized method of moments (GMM) framework. Furthermore, it is worth noting that the parameter in Theorem 1 can also be estimated by simply replacing the expectations with sample averages. Specifically, this involves comparing the average outcome Y and the average treatment D for the instrument values $\widetilde{Z} = 1$ to the average outcome Y and the average treatment D for the instrument values $\widetilde{Z} = 0$, respectively.

It is important to point out that, while increasing the number of instruments decreases the sample size used for estimating the CC-ACR, adding instruments does not inherently increase variance. This is because increasing the number of instruments can increase the share of combined compliers considered, possibly resulting in a variance reduction. For a more detailed discussion, refer to van 't Hoff, Lewbel, and Mellace (2023).

Another noteworthy observation is that the proposed methodology accommodates weak instruments, provided they do not substantially reduce the number of observations and at least one strong instrument is present. As a result, valuable information from weak instruments, such as the 2-year college instrument in Card's 1995 study, can still be obtained.

5.2 Estimation with covariates

Corollary 2 in Section 3.3 provides the identification result of the CC-ACR under the assumption that conditional independence holds, assuming adequate overlap. While it may be tempting to incorporate covariates linearly into the subsample-TSLS approach outlined in the previous section, linear inclusion of the covariates can neglect treatment effect heterogeneity and introduce interpretation complexities already in the context of binary treatments (Blandhol, Bonney, Mogstad & Torgovitsky, 2022; Słoczyński, 2020). Blandhol, Bonney, Mogstad, and Torgovitsky (2022) demonstrate that the linear inclusion of covariates in the TSLS estimator may result in

the inclusion of response types beyond compliers. It is particularly concerning that treatment effects for these additional response types might always receive negative weights.

A fully saturated first stage is necessary to preserve the interpretation of the estimates, which suggests the use of nonparametric estimation methods. With only a few discrete covariates, nonparametric approaches such as kernel regression, local polynomial regression, and series estimators may be suitable (see Frölich, 2007). However, these methods quickly run into the curse of high dimensionality as the number of covariates increases.

To circumvent this issue, I flexibly control for covariates with rich interactions and complex functions using machine learning to handle high-dimensional controls. Specifically, I extend the double debiased machine learning (DML) framework by Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins (2018) to estimate the CC-ACR with covariates of Equation (4), which equals the ratio of two average treatment effects (ATEs), in an interactive IV model.[11] To the best of my knowledge, this is the first paper to use the DML framework in a setting with disrete, ordered or continuous treatments and multiple binary instruments. By focusing on the subsample of size n_s where either $Z_1 = Z_2 = ... = Z_n = 1$ or $Z_1 = Z_2 = ... = Z_n = 0$, and introducing the single binary instrument $\widetilde{Z}$, I reduce the estimation complexity of multiple instruments to a single instrument scenario while maintaining the advantageous interpretation of using multiple instruments. Furthermore, the results do not depend on binary D; only the instrument $\widetilde{Z}$ needs to be binary.

The following interactive instrumental variable model is considered:

$$
\begin{aligned}
Y &= \mu_0(\widetilde{Z}, X) + \nu, \quad E(\nu|\widetilde{Z}, X) = 0, \\
D &= m_0(\widetilde{Z}, X) + U, \quad E(U|\widetilde{Z}, X) = 0, \\
\widetilde{Z} &= p_0(X) + V, \qquad E(V|X) = 0,
\end{aligned}
$$

where ν, U, and V are independent. Define the following functions for the true value of the nuisance parameter $\eta_0 = (\mu_0, m_0, p_0)$:

$$
\begin{aligned}
\mu_0(\widetilde{Z}, X) &= E(Y|\widetilde{Z}, X), \\
m_0(\widetilde{Z}, X) &= E(D|\widetilde{Z}, X), \\
p_0(X) &= E(\widetilde{Z}|X).
\end{aligned}
$$

The nuisance parameter $\eta = (\mu, m, p)$ denotes square-integrable functions μ, m, and p. While μ and m map the support of $(\widetilde{Z}, X)$ to $\mathbb{R}$, p maps the support of X to $(\varepsilon, 1-\varepsilon)$ for some $\varepsilon \in (0, 1/2)$.

Adjusting the methodology of Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins (2018), denote the true parameter of interest $\beta_0^{\text{CC-ACR}}$. The orthogonal score for

[11] For an introduction to DML, see Chernozhukov, Hansen, Kallus, Spindler, and Syrgkanis (2024).

estimating $\beta_0^{\text{CC-ACR}}$ is given by

$$\begin{aligned}
&\psi\left((Y, D, X, \widetilde{Z}); \beta^{\text{CC-ACR}}, \eta\right) \\
&\equiv \mu(1, X) - \mu(0, X) + \left(\frac{\widetilde{Z}}{p(X)} - \frac{(1-\widetilde{Z})}{1-p(X)}\right) \cdot (Y - \mu(\widetilde{Z}, X)) \\
&\quad - \left(m(1, X) - m(0, X) + \left(\frac{\widetilde{Z}}{p(X)} - \frac{(1-\widetilde{Z})}{1-p(X)}\right) \cdot (D - m(\widetilde{Z}, X))\right) \times \beta^{\text{CC-ACR}},
\end{aligned} \tag{10}$$

and at $\eta_0 = (\mu_0, m_0, p_0)$ satisfies the moment condition $E(\psi((Y, D, X, \widetilde{Z}); \beta_0^{\text{CC-ACR}}, \eta_0)) = 0$ and the Neyman orthogonality condition $\partial_\eta E_\psi((Y, D, X, \widetilde{Z}); \beta_0^{\text{CC-ACR}}, \eta_0) = 0$. Then, under similar regularity assumptions as stated in Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins (2018), the behavior of $\hat{\beta}^{\text{CC-ACR}}$ is not affected by the estimation error of the nuisance parameters:

$$\sqrt{n_s}(\hat{\beta}^{\text{CC-ACR}} - \beta_0^{\text{CC-ACR}}) \approx \sqrt{n_s}\mathbb{E}_n(\phi_0(Y, D, X, \widetilde{Z})),$$

where

$$\phi_0(Y, D, X, \widetilde{Z}) = -J_0^{-1}\phi(W; \beta_0^{\text{CC-ACR}}, \eta_0),$$

is the influence function and

$$J_0 := E\left(m_0(1, X) - m_0(0, X)\right),$$

the Jacobian matrix. Subsequent these conditions, $\hat{\beta}^{\text{CC-ACR}}$ is approximately normal:

$$\sqrt{n_s}(\hat{\beta}^{\text{CC-ACR}} - \beta_0^{\text{CC-ACR}}) \overset{a}{\sim} N(0, V), \quad V := E(\phi_0(Y, D, X, \widetilde{Z})\phi_0(Y, D, X, \widetilde{Z})').$$

Estimates of $\hat{\beta}^{\text{CC-ACR}}$ are obtained through plugging in the cross-fitted residuals, constructed by the predictions of the nuisance functions of the machine learners over K folds, into Equation (10). Median estimates over different sample splits can be considered to account for variability in finite samples due to sample splitting, making the estimates more robust to outliers.

Post-regularized inference proceeds as proposed by Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins (2018), with $\hat{V}$ obtained by plugging in the different components and subsequently $\sqrt{\hat{V}/n_s}$ the estimator of the standard error of $\hat{\beta}^{\text{CC-ACR}}$.

6 Empirical applications

The proposed methodology is broadly applicable to various studies involving nonbinary, ordered treatments and multiple instruments, including studies like Attanasio, Maro, and Vera-Hernández (2013). My results further extend to studies using Mendelian randomization, which

employs genetic variants as instruments and has found applications in social sciences and economics (see, for example, Dixon, Hollingworth, Harrison, Davies & Smith, 2020; Scholder, Smith, Lawlor, Propper & Windmeijer, 2013). To provide another example, my results are relevant to applications that instrument for child BMI using the BMI of biological relatives (see, for example, Cawley, 2004; Lindeboom, Lundborg & Van Der Klaauw, 2010; Kline & Tobias, 2008).

To demonstrate the results of the preceding sections, I consider two empirical applications. First, I revisit Card's (1995) seminal study on the returns to education. Second, I apply the novel methodology to Angrist and Evans' (1998) data to study the effect of an additional child on female labor market outcomes.

6.1 The causal effect of schooling on wage

6.1.1 Data

Here, I briefly introduce the most relevant aspects regarding the data set, referring the interested reader to Card (1995) for a more detailed discussion. Card (1995) uses data from the 1979 National Longitudinal Survey of Youth (NLSY79). The data are available in the R package *ivmodel* from Kang, Jiang, Zhao, and Small (2021). The outcome variable is the logarithm of wage, the treatment is years of education ($D \in \{8, 9, ..., 18\}$), and the instruments are indicators for the presence of 2-year and 4-year colleges within a county ($Z_1, Z_2 \in \{0, 1\}$). Given the concerns about instrument validity raised in Card (1995), I adopt a similar approach by incorporating a comprehensive set of controls in the analysis. These controls include dummy variables for race, living in the South, residing in a metropolitan area, and region fixed effects to account for potential systematic variations across different geographical regions. Additionally, IQ is included as a control variable.

Card (1995) primarily focuses on the 4-year college instrument, deeming the 2-year college instrument weak. However, excluding the 2-year college means losing information on men who comply only with this instrument. My methodology can handle weak instruments as long as one is strong. While including the 2-year college reduces observations for CC-ACR estimation by 44% (see Table 4), it adds compliers, potentially offsetting the loss of observations without reducing efficiency (see van 't Hoff, Lewbel, and Mellace (2023)).

6.1.2 Analysis of the causal effect of schooling on wage

In this section, I present the empirical findings from implementing the TSLS methodology as well as the proposed CC-ACR methodology. The results are reported in Table 5 and Figure 1.

Table 4: Number of observations in the full sample and the subsample used for estimating the CC-ACR in the study by Card (1995).

	Nr. obs.	% obs.
Full sample	2,061	100%
Subsample where $\widetilde{Z}$ equals zero or one	1,159	56%

The TSLS specification is saturated in the instruments in the first stage. The TSLS estimates without covariates should be interpreted as the weighted average of the weighted averages in Equation (5) of Proposition 1. It is important to note that differences in estimates can be attributed to differences in underlying complier populations, underlying assumptions, potential violations of the partial monotonicity assumption, or the different weighting schemes. Notably, the confidence intervals of TSLS are comparable to those of CC-ACR, despite CC-ACR using only 56% of the observations. However, it should be noted, that the two methods estimate different causal parameters.

Columns 3 and 4 present the results when using the 2-year and 4-year college instruments individually, while controlling for the other instrument. The estimated effects are imprecise, likely because each instrument lacks sufficient strength. However, when the estimates are combined - shown in Column 6 with the TSLS estimate and Column 8 with the CC-ACR estimate - the estimated effects are significant, leveraging the combined strength of the instruments.

Covariates are included linearly in Columns 6 and 8 of Table 5. This linear inclusion may be problematic in the case of multiple instruments, as it can introduce biases from never-takers and always-takers, potentially contaminating the estimated averages of combined complier effects (see Section 5.2). Ideally, one would include all covariate and instrument interactions to maintain optimal interpretability. However, this fully saturated specification suffers from high dimensionality. Specifically, ignoring the control for IQ for the moment, the number of parameters in a fully saturated first stage with 2 binary instruments and 12 dummies is $2^{14} = 16,384$, which far exceeds the number of observations in the subsample, making it infeasible to use a fully saturated specification.

I now shift my focus to the DML approach discussed in Section 5.2, which offers greater flexibility in accounting for confounding factors, thereby maintaining the interpretation of the CC-ACR. For estimating the nuisance functions, I employ three different machine learning algorithms: Lasso (Tibshirani, 1996), Random Forest (Breiman, 2001), and Boosted Trees (Friedman, 2001). Details regarding specific implementations, including hyperparameter tuning, can be found in Appendix D.1.

The best learner was selected by standardizing the root mean squared error (RMSE) for

Table 5: The causal effect estimates of an additional year of schooling on wage.

	$\hat{\beta}_{\text{OLS}}$		$\hat{\beta}_{\text{2-year}}$	$\hat{\beta}_{\text{4-year}}$	$\hat{\beta}_{\text{TSLS}}$		$\hat{\beta}_{\text{CC-ACR}}$		$\hat{\beta}_{\text{DML-Lasso}}$	$\hat{\beta}_{\text{DML-RF}}$	$\hat{\beta}_{\text{DML-Boosted}}$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Years of schooling	0.039***	0.027***	0.286	0.072	0.254***	0.117*	0.268***	0.170*	0.147***	0.122***	0.136**
(Std. err.)	(0.004)	(0.004)	(0.217)	(0.069)	(0.060)	(0.064)	(0.068)	(0.070)	(0.047)	(0.042)	(0.053)
Observations	2,061	2,061	2,061	2,061	2,061	2,061	1,159	1,159	1,159	1,159	1,159
% observations	100%	100%	100%	100%	100%	100%	56%	56%	56%	56%	56%
Covariates	no	linear	linear	linear	no	linear	no	linear	flexible	flexible	flexible

Note. This table presents the estimates of causal effects of an additional year of schooling on wage for different causal parameters and estimation approaches. $\hat{\beta}_{\text{2-year}}$ and $\hat{\beta}_{\text{4-year}}$ give the instrument-specific LATE when using the instruments separately. For columns (9), (10), and (11), results are obtained using five-fold cross-fitting. For columns (9), (10), and (11), median estimates and standard errors across 25 splits are reported to take into account different sample splits.

*Significance level: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

predicting the three nuisance functions and choosing the learner that minimized the sum of these errors (see Table 13 in Appendix D.1). The estimate using the best-performing learner (Boosted Trees) is presented in Figure 1, though the mean RMSE values across all machine learners are similar (see Table 13 in Appendix D.1), as well as the CC-ACR estimates, indicating that the primary insights are unaffected by the choice of learner.

The CC-ACR estimate with the best learner implies that a one-year increase in education is associated with an average wage increase of approximately 14% for the combined complier population. This is slightly higher than the estimate found with TSLS, though it should be noted that TSLS requires imposing PM instead of LiM, and PM might be violated, leading to an undesirable interpretation. The large effect might be attributed to the inclusion of the 2-year college, which adds compliers likely at the lower end of the wage distribution. The CC-ACR estimate therefore includes individuals who might benefit more from schooling than those affected by the presence of a 4-year college or those unaffected by the presence of a college, aligning with earlier arguments theorized by Card (1995). This is particularly interesting because it provides insights about the 2-year college compliers, even though the 2-year college instrument is a weak instrument and does not provide consistent estimates by itself. Adding 2-year college compliers is possible thanks to the strong 4-year college instrument, meaning that CC-ACR using both instruments offers additional insights.

Finally, the CC-ACR estimates for the years of schooling variable account for individuals who attended some college but did not complete a degree, in contrast to the binary college diploma variable, which only indicates the absence of a degree. Exploring how these differences in estimates could support either human capital theory (education enhances productivity) or signaling theory (education signals ability to employers) would be valuable but is beyond the scope of this study.

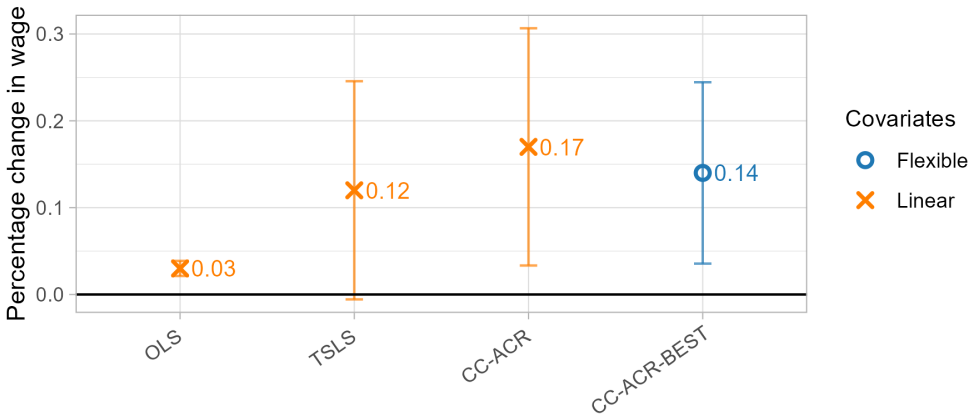


Figure 1: This figure presents some of the estimates of the causal effect of schooling on wage reported in Table 5 for easier comparison. CC-ACR-BEST is the DML estimate obtained using the machine learner that achieved the lowest sum of standardized mean RMSE

6.1.3 Weighting function of the CC-ACR

As shown in Section 3.6, the unstandardized weighting function can be obtained by subtracting the CDFs of the treatment conditional on the instrument values as follows: $P(D < j|Z_1 = Z_2 = ... = Z_K = 0)$ and $P(D < j|Z_1 = Z_2 = ... = Z_K = 1)$. Figure 2 depicts the two CDFs and Figure 3 depicts their unstandardized difference. Standardizing this to sum to one (that is, normalizing them by the first stage), this function provides the weights associated with the per-level effects along the causal response function of the CC-ACR as presented by Equation (6). These weights reflect the combined strength of the instruments and are informative about the complier distribution across the range of treatment values.

For each year of schooling j, the difference is the share of individuals whose education increases from less than j years to j years or more in response to the shift in instrument values. Unsurprisingly, the figure shows that more weight is concentrated around 12 to 14 years of education. This period, just after high school, is when most individuals make decisions about attending college, and thus, when the instruments (such as the presence of colleges) are most likely to influence educational attainment. This observation aligns with the findings of Kling (2001). While it is not possible to determine the exact size of the complier population due to potential overlaps at different treatment values, Figure 3 offers an estimate of a lower bound. The data indicates that at least 13% (the maximum value of shares) belongs to the combined complier population.

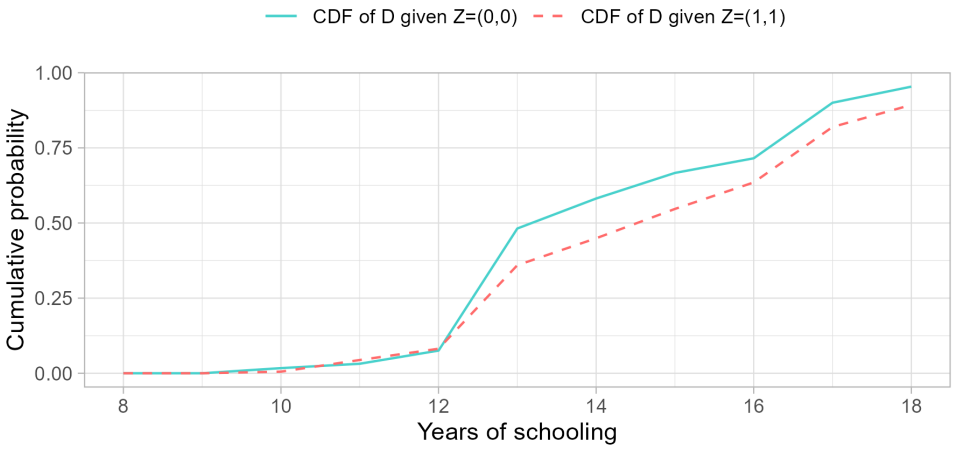


Figure 2: CDFs of the treatment conditional on the outer support of the instrument values in Card's (1995) application. When the CDFs cross, this might provide a visual indication that the necessary condition for LiM does not hold at that treatment level.

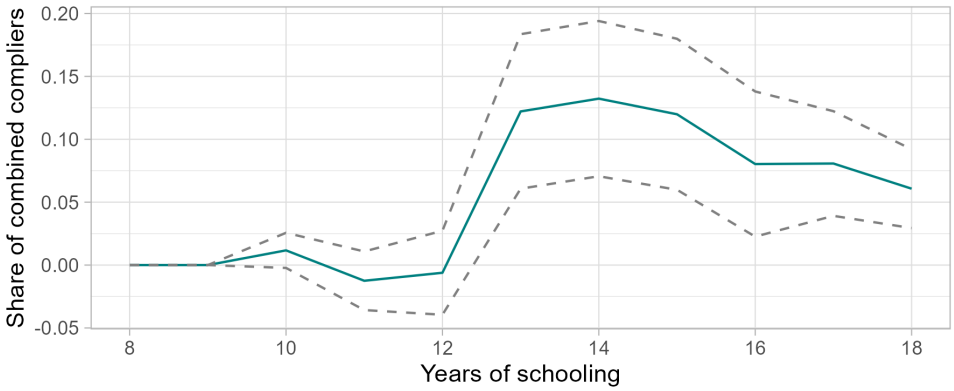


Figure 3: The unstandardized weighting function of the CC-ACR parameter of Equation (6) in Card's (1995) application. Standardizing these values to sum up to one gives the weights on the average causal responses of the combined complier types. 95% confidence intervals are calculated in a standard fashion for a difference in proportions and indicated by the dashed lines.

Table 6: Number of initial response types for different combinations of treatment and instrument values.

Treatment	Instrument(s)	Nr. Response Types
$D \in \{0, 1\}$	$Z_1 \in \{0, 1\}$	4
$D \in \{0, 1\}$	$Z_1, Z_2 \in \{0, 1\}$	16
$D \in \{0, 1, 2\}$	$Z_1, Z_2 \in \{0, 1\}$	81
$D \in \{8, 9, ..., 18\}$	$Z_1, Z_2 \in \{0, 1\}$	14,641

Table 7: Two example types in Card's (1995) study. PM rules out one of these two types, imposing that individuals either prefer the 2-year college or the 4-year college. Both types are consistent with LiM ($P(D^{11} \geq D^{00}) = 1$). Z_1 is the instrument for the proximity of a 2-year college, and Z_2 for the proximity of a 4-year college.

Type	D^{00}	D^{01}	D^{10}	D^{11}
$t_{12,14,12,14}$	12	14	12	14
$t_{12,14,16,14}$	12	14	16	14

6.1.4 Plausibility of LiM

In Card's 1995 study, there are $(J + 1)^{2^K} = 11^4 = 14,641$ different initial response types (see Table 6). Identifying a causal effect necessitates ruling out some types by imposing monotonicity. LiM is arguably more plausible than PM in this context.

To illustrate, consider the two types in Table 7. The first type, $t_{12,14,12,14}$, does not attend college when no college is nearby ($D^{00} = 12$), attends a 2-year college if one is nearby ($D^{01} = 14$), but does not respond to the presence of a 4-year college ($D^{10} = 12$). The second type, $t_{12,14,16,14}$, does not attend college when no college is nearby ($D^{00} = 12$), attends a 2-year college if one is nearby ($D^{01} = 14$), and attends a 4-year college if one is nearby ($D^{10} = 16$). Both types are likely to exist, but PM would rule out one, as it requires either $P(D^{10} \geq D^{11}) = 1$ or $P(D^{10} \leq D^{11}) = 1$. While PM requires weak monotonicity in the propensity score function with respect to instruments, LiM does not, and thus allows for the coexistence of the two example types in Table 7 and many more.

Interestingly, Mogstad, Torgovitsky, and Walters (2021) highlight that a key limitation of IAM is its inherent preference for one instrument over another, which restricts choice behavior. However, when extending PM to the nonbinary treatment setting, PM faces similar restrictions.

6.1.5 LiM test

LiM can be violated if there exist defiers with respect to one instrument who cannot be convinced to comply by another instrument. Consider the case of defiers with respect to the 4-year college instrument. These individuals may be negatively influenced by proximity to a 4-year college due to negative interactions with the college community, unfavorable perceptions of the institution's culture, or discouraging information about attending college. Furthermore, familiarity with the nearby college might diminish its perceived prestige, deterring them from pursuing a 4-year education. If such defiers cannot be persuaded to comply by the presence of a 2-year college, LiM is violated.

A visual inspection of Figure 2 hints at a potential violation of LiM for low treatment values, as the CDFs cross when the number of years if schooling is below 12 years. To address potential violations of LiM within subgroups of observed characteristics, I implement the local LiM test as described in Section 4.2. For this, I adapt the R package *LATEtest* developed by Farbmacher, Guber, and Klaassen (2022).[12] The test indicates no detected violation of the LiM assumption.

6.2 The causal effect of an additional child on female labor market outcomes

In this section, I apply the proposed methodology to the data from Angrist and Evans (1998) to analyze the impact of an additional child on female labor market outcomes.

6.2.1 Data

The data for the analysis is derived from the 1980 Census Public Use Micro Data Samples (PUMS). I consider the sample of married women. For an in-depth discussion of the data, see Angrist and Evans (1998). The primary outcome variables considered are a mother's annual labor income, the hours worked per week, and the weeks worked per year. The treatment variable is the number of children ($D \in \{2, ..., 6\}$). Families with more than six children are excluded to avoid low representation when sample splitting for the DML-based estimation approach, to aid in the interpretation of the estimated CC-ACR, and since the data only contains information for the first five children born.

I consider two instruments for the analysis. The first instrument, Z_1, is a binary variable that equals one if there are twins at the second or subsequent births and zero otherwise. To the best of my knowledge, this instrument has not been previously used in the literature. Unlike the twinning instrument by Angrist and Evans (1998), which is specific to the birth of a third child, this novel instrument provides incentives for having an additional child for any number

[12] The procedure's configurations are as follows: The fraction of data used for each tree equals its default value of 0.5, and the minimum size of control and treated observations per leaf is set to 50.

of children between 2 and 6. The second instrument, Z_2, is the classical same-sex instrument, which is equal to one when the first two children are of the same sex.[13]

Following Angrist and Evans (1998), the covariates included in the analysis are mother's age, age at first birth, sex of the first-born, and indicators for race and Hispanic ethnicity. Additionally, the analysis controls for the sex of the second-born, the marital status of the mothers, and the mothers' highest level of educational attainment.

6.2.2 Analysis of the causal effect of an additional child on female labor market outcomes

This section presents the causal effects of an additional child on three female labor market outcomes: annual labor income, hours worked per week, and weeks worked per year, using different estimation techniques. Table 8 provides estimates, some of which are emphasized in corresponding Figure 4.

When using the novel twinning instrument that considers twinning at any birth (Column 3), no significant effect is observed. This contrasts with the original paper, which focused on twinning at the second birth. One possible explanation for this difference is that women who have an additional child at, for example, the fourth birth might already be inclined to have larger families and be less active on the labor market. As a result, compliers with respect to the novel twinning instrument likely experience a smaller effect of an additional child on their labor market outcomes.

The TSLS and CC-ACR estimates combining both instruments and linearly including covariates, presented in Columns 6 and 8, respectively, indicate no effect of an additional child on annual labor income (Panel A), and only the TSLS estimate indicates a reduction of 0.896 hours worked per week. Columns 9 and 11 display the results of the CC-ACR estimates obtained using DML, which flexibly accounts for covariates. Detailed specifications for the machine learning models are provided in Appendix D.2. Tree-based methods reveal a significant, albeit small, reduction in labor income between \$97.245 and \$70.832 (Panel A), with the latter estimate produced by using Boosted Trees as learners for the nuisance parameters, which achieved the minimum combined RMSE. In contrast, with Lasso as learner, the effect estimates are not significant. This may be attributed to only including third-order interactions (see Appendix D.2), which suggests that the poor performance of Lasso could be due to the highly complex nature of the relationships involved. Across the machine learning methods, a significant reduction in hours worked per week is observed, ranging from 0.59 to 0.895 hours (Panel B). Additionally,

[13]Note that using this instrument is equivalent to using an instrument that equals one if the first two, three, four, or more children are all of the same sex and zero otherwise.

Table 8: The causal effect estimates of an additional child on female labor market outcomes.

	$\hat{\beta}_{\text{OLS}}$		$\hat{\beta}_{\text{twinning}}$	$\hat{\beta}_{\text{same-sex}}$	$\hat{\beta}_{\text{TSLS}}$		$\hat{\beta}_{\text{CC-ACR}}$		$\hat{\beta}_{\text{DML-Lasso}}$	$\hat{\beta}_{\text{DML-RF}}$	$\hat{\beta}_{\text{DML-Boosted}}$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Panel A: Causal effect on labor income											
Additional child	−709.900***	−1022.451***	44.864	−631.253***	-143.230	-81.720	40.500	191.718	19.016	−97.245***	−70.832***
(Std. err.)	(13.930)	(14.203)	(93.878)	(229.681)	(95.140)	(86.734)	(14.200)	(124.279)	(20.723)	(23.309)	(19.574)
Observations	207,674	207,674	207,674	207,674	207,674	207,674	103,935	103,935	103,935	103,935	103,935
% observations	100%	100%	100%	100%	100%	100%	50%	50%	50%	50%	50%
Covariates	no	linear	linear	linear	no	linear	no	linear	flexible	flexible	flexible
Panel B: Causal effect on hours worked per week											
Additional child	−2.468***	−4.137***	-0.167	−4.055***	-1.780	−0.896**	−1.167*	0.340	−0.895***	−0.873***	−0.590***
(Std. err.)	(0.054)	(0.056)	(0.372)	(0.981)	(0.367)	(0.339)	(0.552)	(0.487)	(0.080)	(0.091)	(0.076)
Observations	207,674	207,674	207,674	207,674	207,674	207,674	103,935	103,935	103,935	103,935	103,935
% observations	100%	100%	100%	100%	100%	100%	50%	50%	50%	50%	50%
Covariates	no	linear	linear	linear	no	linear	no	linear	flexible	flexible	flexible
Panel C: Causal effect on weeks worked											
Additional child	−3.590***	−5.554***	-0.316	−5.033***	−2.141***	−1.160**	−1.278*	0.452	−0.944***	−0.940***	−0.605***
(Std. err.)	(0.065)	(0.066)	(0.444)	(1.166)	(0.441)	(0.404)	(0.660)	(0.582)	(0.096)	(0.108)	(0.091)
Observations	207,674	207,674	207,674	207,674	207,674	207,674	103,935	103,935	103,935	103,935	103,935
% observations	100%	100%	100%	100%	100%	100%	50%	50%	50%	50%	50%
Covariates	no	linear	linear	linear	no	linear	no	linear	flexible	flexible	flexible

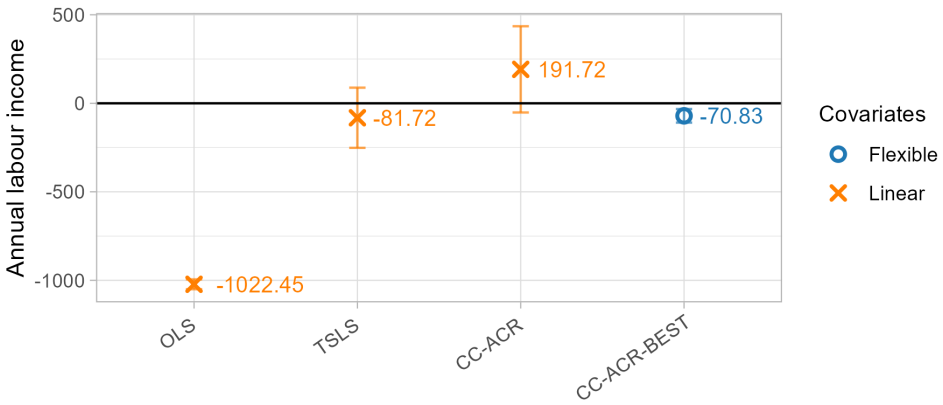
Note. This table presents the estimates of causal effects of an additional child on female labor market outcomes for different causal parameters and estimation approaches.. $\hat{\beta}_{\text{twinning}}$ and $\hat{\beta}_{\text{same-sex}}$ give the instrument-specific LATE when using the instruments separately. For columns (9), (10), and (11), results are obtained using five-fold cross-fitting. For columns (9), (10), and (11), median estimates and standard errors across 5 splits are reported to take into account different sample splits.

*Significance level: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

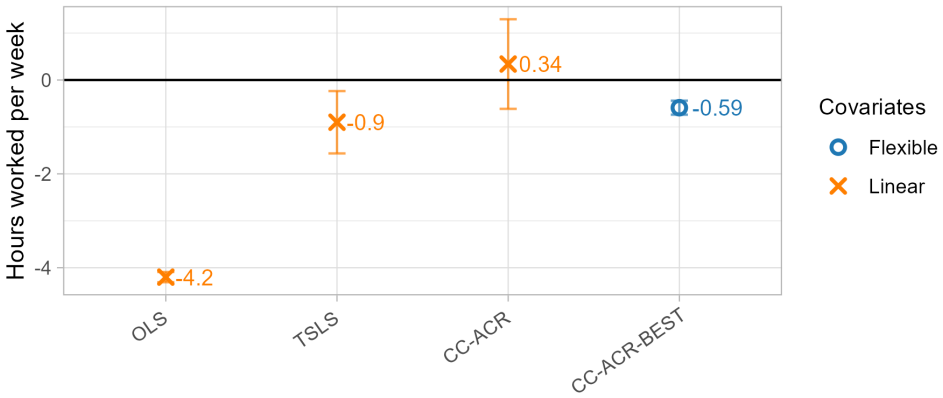
weeks worked decreased significantly, with reductions between 0.61 and 0.94 weeks (Panel B). Again, Boosted Trees outperform the other learners in terms of the RMSE (see Appendix D.2) and the effect estimates indicate a significant reduction of 0.59 hours worked per week and 0.61 weeks worked.

The results presented in this section complement those of Angrist and Evans (1998). The CC-ACR estimates the average effect of having any additional child after the first, which differs from the original study by Angrist and Evans (1998) that focuses on the effect of having more than two children. Studies using the twinning instrument at the second birth or the same-sex of the first two children are restricted to estimating the impact of a third child. In contrast, the CC-ACR approach captures the effect of any additional child after the first, as the novel twinning instrument affects the likelihood of having a second child or more, while the same-sex instrument influences the birth of a third child or more. Second, Angrist and Evans (1998) employ each instrument individually. A key advantage of using multiple instruments, as done in this study, is that it accommodates defiers with respect to the same-sex instrument. These defiers do not affect the CC-ACR estimates as long as they can be induced to comply through the influence of the twinning instrument.

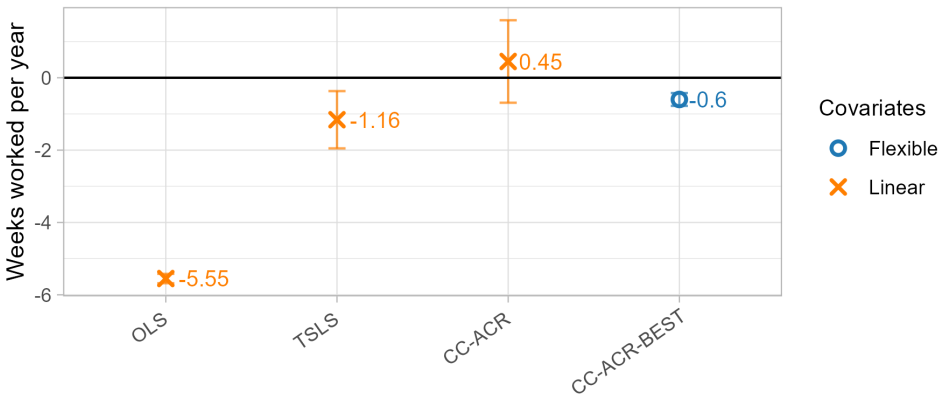
With respect to annual labor income, Angrist and Evans (1998) report reductions of $1,338



(a) Causal effect of an additional child on annual labor income.



(b) Causal effect of an additional child on hours worked per week.



(c) Causal effect of an additional child on weeks worked.

Figure 4: This figure presents some of the estimates reported in Table 8 for easier comparison. CC-ACR-BEST is the DML estimate obtained using the machine learner that achieved the lowest sum of standardized mean RMSE.

Table 9: Two example types in the Angrist and Evans (1998) application. PM rules out one of these two types. Both types are consistent with LiM ($P(D^{11} \geq D^{00}) = 1$). Z_1 is the twinning instrument, and Z_2 the same-sex instrument.

Type	D^{00}	D^{01}	D^{10}	D^{11}
$t_{3,2,3,3}$	3	2	3	3
$t_{2,3,3,3}$	2	3	3	3

using the same-sex instrument and $1,308 using the twinning instrument for married women, specifically for the effect of having a third child. For hours worked per week, they find reductions of 4.87 and 4.59 hours, respectively. For weeks worked, they find reductions of 5.45 and 5.15 weeks, respectively. When compared to the CC-ACR estimates in Column 11 of Table 8, the negative effect of an additional child, when considering up to six children, is noticeably smaller than the effect of a third child as estimated by Angrist and Evans (1998). This discrepancy likely arises because women who, for instance, have twins at the fourth or fifth birth may have different labor market preferences and are already predisposed to lower labor market participation, thereby attenuating the observed effect.

6.2.3 Weighting function of the CC-ACR

The CC-ACR weighting function provides insights into the shares of women at different household sizes who have more children because of the instrument incentives. Figure 6 shows an absence of compliers at the treatment level of two children, which is expected as the instruments incentivize having three or more children. Additionally, approximately 35% of the population are combined compliers, as depicted in Figure 6. This means that at least 35% of the women have an additional child because of twins or same-sex preferences.

6.2.4 Plausibility of LiM

Consider the two example types in Table 9. A woman of type $t_{3,2,3,3}$ has a strong preference for two girls: she has three children if the first two are a boy and a girl ($D^{00} = 3$), two children if the first two are girls ($D^{01} = 2$), and three children if twins are born at the second birth ($D^{10} = 3$ and $D^{11} = 3$). Another woman, $t_{2,3,3,3}$, prefers the first two children to be of different sexes: she has two children if the first two are of mixed sex ($D^{00} = 2$), and three children if the first two are of the same sex ($D^{01} = 3$). PM does not hold here because it requires either $P(D^{10} \geq D^{00}) = 1$ or $P(D^{10} \leq D^{00}) = 1$.

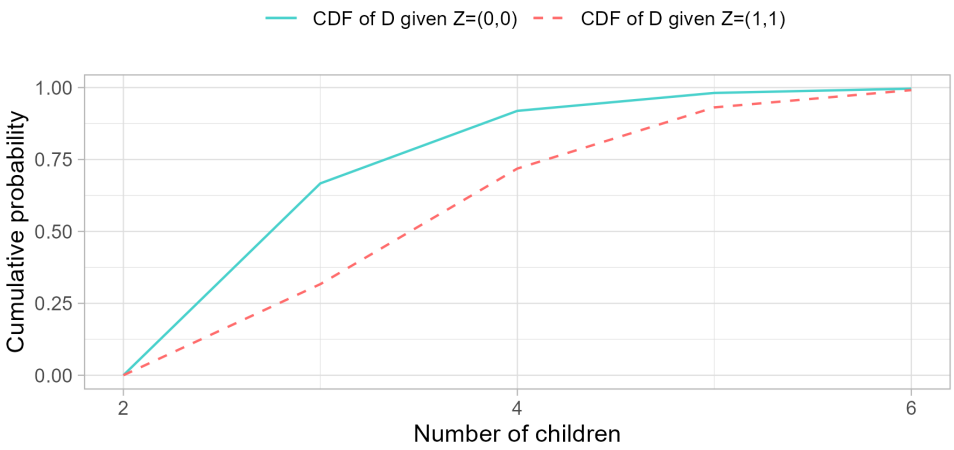


Figure 5: CDFs of the treatment conditional on the outer support of the instrument values in Angrist and Evans' (1998) application. The CDFs do not cross, indicating that the necessary condition for LiM holds at all treatment levels and no violation of the LiM assumption is detected visually in the full sample.

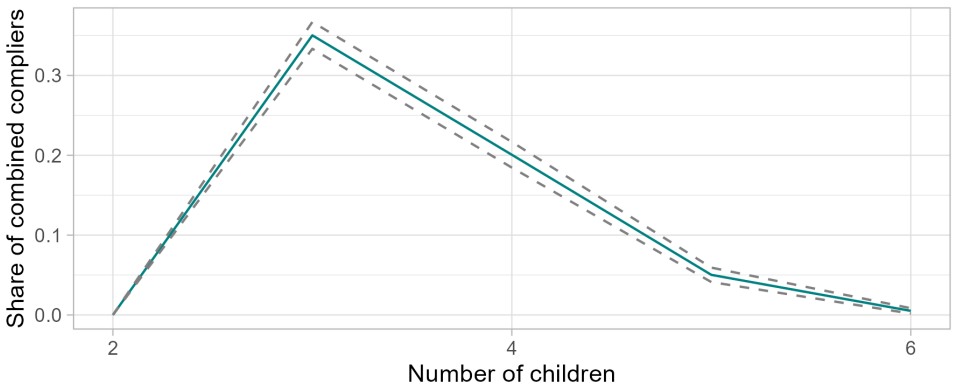


Figure 6: The unstandardized weighting function of the CC-ACR parameter of Equation (6) in Angrist and Evans' (1998) application. Standardizing these values to sum up to one gives the weights on the average causal responses of the combined complier types. 95% confidence intervals are calculated in a standard fashion for a difference in proportions and indicated by the dashed lines.

6.2.5 LiM test

In the study by Angrist and Evans (1998), it is very unlikely that LiM is violated, as one cannot defy the twinning instrument, which always pushes towards compliance (i.e., having an additional child).[14] Consistent with this, Figure 5 visually provides no evidence of a LiM violation. Unsurprisingly, the local LiM test (see Section 4.2) does not detect a violation.[15]

7 Conclusion

This study advances the understanding of causal inference in the context of discrete, ordered and continuous treatments with multiple instruments by introducing a novel approach to identification and estimation. The central theoretical contribution, the CC-ACR, provides an intuitive and robust alternative to the conventional TSLS methodology. Unlike TSLS, the CC-ACR accommodates a less restrictive LiM assumption and offers a clearer interpretation of causal effects by leveraging a weighting scheme directly tied to the shares of combined complier types. This approach is particularly valuable for estimating the average causal effect of a one-level increase in treatment.

The study also introduces a practical test for the LiM assumption, using causal forests to detect local violations in a data-driven manner. This test enhances the robustness of empirical findings by identifying potential subgroups where the LiM assumption may not hold.

Empirical applications underscore the practical advantages of the CC-ACR framework under the LiM assumption. In estimating the returns to education, the CC-ACR allows for more granular insights into the impact of each additional year of schooling on wages, surpassing the binary approach that only considers college attendance. By leveraging both strong and weak instruments, this method delivers more precise estimates and policy-relevant insights. Similarly, in the analysis of the effect of additional children on female labor market outcomes, the CC-ACR indicates heterogeneity in labor market preferences, facilitated by incorporating the novel twinning instrument. The CC-ACR estimates are obtained by adapting the DML methodology of Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins (2018) to the setting with discrete, ordered and continuous treatments and multiple instruments.

The present study also reveals several promising avenues for future research. For instance, it is straightforward to extend the results to fuzzy regression discontinuity designs with a multivalued treatment and multiple running variables. Additionally, future research could assess potential power improvements for the LiM test. Techniques such as pruning or replacing the Bonferroni correction with a multiplier bootstrap approach, in the spirit of Huber and Kueck

[14] Note that van 't Hoff, Lewbel, and Mellace (2023) provide a related discussion in case of a binary treatment.

[15] Configuration settings for the LiM test are the same as those specified in footnote 12.

(2022), hold potential. Moreover, while the CC-ACR represents a weighted average of causal effects, one might instead be interested in obtaining the causal effect of a one-level increase, $(Y^j - Y^{j-1})$, for some specific treatment level $j \in \{1, ..., J\}$. Building upon the work of Kitagawa (2021) and Huber, Laffers, and Mellace (2017), partial identification could be explored for the average causal response resulting from, for example, a one-level increase in the treatment level for various combined complier groups. Note that for point-identifying the effect along specific treatment margins for a combined complier population, an instrument that pushes individuals towards compliance at that specific margin is required (see, for instance, Bhuller and Sigstad, 2022).

References

Abadie, A. (2002). Bootstrap Tests for Distributional Treatment Effects in Instrumental Variable Models. *Journal of the American Statistical Association*, *97*(457), 284–292 (cit. on p. 19).

Angrist, J. D., & Evans, W. N. (1998). Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size. *American Economic Review*, *88*(3), 450–477 (cit. on pp. 3, 4, 24, 30–32, 34–36, 59).

Angrist, J. D., Graddy, K., & Imbens, G. W. (2000). The Interpretation of Instrumental Variables Estimators in Simultaneous Equations Models with an Application to the Demand for Fish. *The Review of Economic Studies*, *67*(3), 499–527 (cit. on p. 52).

Angrist, J. D., & Imbens, G. W. (1995). Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity. *Journal of the American Statistical Association*, *90*(430), 431–442 (cit. on pp. 1, 2, 4–6, 15–17, 45, 49).

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of Causal Effects Using Instrumental Variables. *Journal of the American Statistical Association*, *91*(434), 444–455 (cit. on pp. 4, 20).

Attanasio, O. P., Maro, V. D., & Vera-Hernández, M. (2013). Community Nurseries and the Nutritional Status of Poor Children. Evidence from Colombia. *The Economic Journal*, *123*(571), 1025–1058 (cit. on p. 23).

Balke, A., & Pearl, J. (1997). Bounds on Treatment Effects from Studies with Imperfect Compliance. *Journal of the American Statistical Association*, *92*(439), 1171–1176 (cit. on p. 5).

Bhuller, M., & Sigstad, H. (2022). 2SLS with Multiple Treatments. *arXiv preprint arXiv:2205.07836* (cit. on pp. 5, 37).

Blandhol, C., Bonney, J., Mogstad, M., & Torgovitsky, A. (2022). *When is TSLS Actually LATE?* (Working Paper No. w29709). National Bureau of Economic Research. Cambridge, MA. (Cit. on p. 21).

Borusyak, K., & Jaravel, X. (2018). *Revisiting Event Study Designs* (SSRN Scholarly Paper No. 2826228). (Cit. on p. 1).

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*, 5–32 (cit. on pp. 25, 58).

Card, D. (1995). Using Geographic Variation in College Proximity to Estimate the Return to Schooling. *Aspects of Labor Market Behaviour: Essays in Honour of John Vanderkamp*, 201–222 (cit. on pp. 3, 4, 6, 7, 11, 21, 24–26, 28, 29, 56, 59).

Carneiro, P., Heckman, J. J., & Vytlacil, E. J. (2011). Estimating Marginal Returns to Education. *American Economic Review*, *101*(6), 2754–2781 (cit. on p. 1).

Carr, T., & Kitagawa, T. (2021). Testing Instrument Validity with Covariates. *arXiv preprint arXiv:2112.08092* (cit. on p. 5).

Cawley, J. (2004). The Impact of Obesity on Wages. *Journal of Human resources*, *39*(2), 451–474 (cit. on p. 24).

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/Debiased Machine Learning for Treatment and Structural Parameters. *The Econometrics Journal*, *21*(1), C1–C68 (cit. on pp. 3, 22, 23, 36, 54).

Chernozhukov, V., Chetverikov, D., Kato, K., & Koike, Y. (2023). High-Dimensional Data Bootstrap. *Annual Review of Statistics and Its Application*, *10*, 427–449 (cit. on p. 20).

Chernozhukov, V., Hansen, C., Kallus, N., Spindler, M., & Syrgkanis, V. (2024). Applied Causal Inference Powered by ML and AI. *rem*, *12*(1), 338 (cit. on p. 22).

De Chaisemartin, C., & d'Haultfoeuille, X. (2020). Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects. *American Economic Review*, *110*(9), 2964–2996 (cit. on p. 1).

Dixon, P., Hollingworth, W., Harrison, S., Davies, N. M., & Smith, G. D. (2020). Mendelian Randomization Analysis of the Causal Effect of Adiposity on Hospital Costs. *Journal of Health Economics*, *70*, 102300 (cit. on p. 24).

Farbmacher, H., Guber, R., & Klaassen, S. (2022). Instrument Validity Tests with Causal Forests. *Journal of Business & Economic Statistics*, *40*(2), 605–614 (cit. on pp. 2, 5, 20, 21, 30, 53–55).

Frandsen, B., Lefgren, L., & Leslie, E. (2023). Judging Judge Fixed Effects. *American Economic Review*, *113*(1), 253–277 (cit. on p. 5).

Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *Annals of Statistics*, 1189–1232 (cit. on pp. 25, 58).

Frölich, M. (2007). Nonparametric IV Estimation of Local Average Treatment Effects with Covariates. *Journal of Econometrics*, *139*(1), 35–75 (cit. on pp. 3, 4, 7, 12, 16, 22).

Goff, L. (2024). A Vector Monotonicity Assumption for Multiple Instruments. *Journal of Econometrics*, *241*(1), 105735 (cit. on pp. 4, 7).

Goodman-Bacon, A. (2021). Difference-in-Differences with Variation in Treatment Timing. *Journal of Econometrics*, *225*(2), 254–277 (cit. on p. 1).

Heckman, J. J., Urzua, S., & Vytlacil, E. (2006). Understanding Instrumental Variables in Models with Essential Heterogeneity. *The Review of Economics and Statistics*, *88*(3), 389–432 (cit. on p. 5).

Huber, M., & Kueck, J. (2022). Testing the Identification of Causal Effects in Observational Data. *arXiv preprint arXiv:2203.15890* (cit. on pp. 20, 36, 37).

Huber, M., Laffers, L., & Mellace, G. (2017). Sharp IV Bounds on Average Treatment Effects on the Treated and Other Populations under Endogeneity and Noncompliance. *Journal of Applied Econometrics*, *32*(1), 56–79 (cit. on p. 37).

Huber, M., & Mellace, G. (2015). Testing Instrument Validity for LATE Identification based on Inequality Moment Constraints. *Review of Economics and Statistics*, *97*(2), 398–411 (cit. on p. 5).

Imbens, G., & Angrist, J. (1994). Identification and Estimation of Local Average Treatment Effects. *Econometrica*, *62*(2), 467–476 (cit. on pp. 4, 7).

Kang, H., Jiang, Y., Zhao, Q., & Small, D. S. (2021). Ivmodel: An R Package for Inference and Sensitivity Analysis of Instrumental Variables Models with One Endogenous Variable. *Observational Studies*, *7*(2), 1–24 (cit. on p. 24).

Kitagawa, T. (2015). A Test for Instrument Validity. *Econometrica*, *83*(5), 2043–2063 (cit. on p. 5).

Kitagawa, T. (2021). The Identification Region of the Potential Outcome Distributions Under Instrument Independence. *Journal of Econometrics*, *225*(2), 231–253 (cit. on pp. 5, 37).

Klein, T. J. (2010). Heterogeneous Treatment Effects: Instrumental Variables Without Monotonicity? *Journal of Econometrics*, *155*(2), 99–116 (cit. on p. 20).

Kline, B., & Tobias, J. L. (2008). The Wages of BMI: Bayesian Analysis of a Skewed Treatment–Response Model with Nonparametric Endogeneity. *Journal of Applied Econometrics*, *23*(6), 767–793 (cit. on p. 24).

Kling, J. R. (2001). Interpreting Instrumental Variables Estimates of the Returns to Schooling. *Journal of Business & Economic Statistics*, *19*(3), 358–364 (cit. on p. 27).

Lee, S., & Salanié, B. (2018). Identifying Effects of Multivalued Treatments. *Econometrica*, *86*(6), 1939–1963 (cit. on p. 5).

Lindeboom, M., Lundborg, P., & Van Der Klaauw, B. (2010). Assessing the Impact of Obesity on Labor Market Outcomes. *Economics & Human Biology*, *8*(3), 309–319 (cit. on p. 24).

Mogstad, M., & Torgovitsky, A. (2024). *Instrumental Variables with Unobserved Heterogeneity in Treatment Effects* (tech. rep.). National Bureau of Economic Research. (Cit. on p. 5).

Mogstad, M., Torgovitsky, A., & Walters, C. R. (2021). The Causal Interpretation of Two-Stage Least Squares with Multiple Instrumental Variables. *American Economic Review*, *111*(11), 3663–98 (cit. on pp. 1, 2, 4, 7, 13, 29).

Mourifié, I., & Wan, Y. (2017). Testing Local Average Treatment Effect Assumptions. *Review of Economics and Statistics*, *99*(2), 305–313 (cit. on p. 5).

Noack, C. (2021). Sensitivity of LATE Estimates to Violations of the Monotonicity Assumption. *arXiv preprint arXiv:2106.06421* (cit. on p. 20).

Robins, J. (1986). A New Approach to Causal Inference in Mortality Studies with a Sustained Exposure Period—Application to Control of the Healthy Worker Survivor Effect. *Mathematical Modelling*, *7*(9), 1393–1512 (cit. on p. 6).

Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1994). Estimation of Regression Coefficients When Some Regressors are not Always Observed. *Journal of the American Statistical Association*, *89*(427), 846–866 (cit. on p. 53).

Rubin, D. B. (1974). Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology*, *66*(5), 688 (cit. on p. 6).

Scholder, S. v. H. K., Smith, G. D., Lawlor, D. A., Propper, C., & Windmeijer, F. (2013). Child Height, Health and Human Capital: Evidence Using Genetic Markers. *European Economic Review*, *57*, 1–22 (cit. on p. 24).

Słoczyński, T. (2020). When Should We (Not) Interpret Linear IV Estimands as LATE? *arXiv preprint arXiv:2011.06695* (cit. on p. 21).

Sun, Z. (2023). Instrument Validity for Heterogeneous Causal Effects. *Journal of Econometrics*, *237*(2), 105523 (cit. on p. 5).

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *58*(1), 267–288 (cit. on pp. 25, 58).

van 't Hoff, N., Lewbel, A., & Mellace, G. (2023). *Limited Monotonicity and the Combined Compliers LATE* (Boston College Working Papers in Economics No. 1059). Boston College. (Cit. on pp. 2, 4, 6, 8, 21, 24, 36).

Wager, S., & Athey, S. (2018). Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests. *Journal of the American Statistical Association*, *113*(523), 1228–1242 (cit. on p. 20).

Appendices

A Complete table with response types

Table 10: Complete table with response types in case of three-valued treatment, $D \in \{0, 1, 2\}$, and two instruments, Z_1 and Z_2. Suppose that the instrument support $\mathcal{Z} = \{z_0, z_1, z_2, z_3\}$ is ordered such that $E(D|Z = z_0) < E(D|Z = z_1) < E(D|Z = z_2) < E(D|Z = z_3)$ and label the ordered elements as z_0, z_1, z_2, z_3. Here, suppose $\mathcal{Z} = \{z_0, z_1, z_2, z_3\} = \{(0, 0), (0, 1), (1, 0), (1, 1)\}$. ✓indicates the response types under the different forms of the monotonicity assumption.

		D^{z_0}	D^{z_1}	D^{z_2}	D^{z_3}			
Combined type	Type	D^{00}	D^{01}	D^{10}	D^{11}	LiM	PM	IAM
$cn_{2,2}$	$n_{2,2,2,2}$	2	2	2	2	✓	✓	✓
	$n_{2,1,2,2}$	2	1	2	2	✓		
	$n_{2,0,2,2}$	2	0	2	2	✓		
	$n_{2,2,1,2}$	2	2	1	2	✓		
	$n_{2,1,1,2}$	2	1	1	2	✓		
	$n_{2,0,1,2}$	2	0	1	2	✓		
	$n_{2,2,0,2}$	2	2	0	2	✓		
	$n_{2,1,0,2}$	2	1	0	2	✓		
	$n_{2,0,0,2}$	2	0	0	2	✓		
$cc_{1,2}$	$c_{1,2,2,2}$	1	2	2	2	✓	✓	✓
	$c_{1,1,2,2}$	1	1	2	2	✓	✓	✓
	$c_{1,0,2,2}$	1	0	2	2	✓		
	$c_{1,2,1,2}$	1	2	1	2	✓	✓	
	$c_{1,1,1,2}$	1	1	1	2	✓	✓	✓
	$c_{1,0,1,2}$	1	0	1	2	✓		
	$c_{1,2,0,2}$	1	2	0	2	✓		
	$c_{1,1,0,2}$	1	1	0	2	✓		
	$c_{1,0,0,2}$	1	0	0	2	✓		
$cc_{0,2}$	$c_{0,2,2,2}$	0	2	2	2	✓	✓	✓
	$c_{0,1,2,2}$	0	1	2	2	✓	✓	✓
	$c_{0,0,2,2}$	0	0	2	2	✓	✓	✓
	$c_{0,2,1,2}$	0	2	1	2	✓	✓	

Continued on next page

Table 10 – continued from previous page

Combined type	Type	D^{00}	D^{10}	D^{01}	D^{11}	LiM	PM	IAM
	$c_{0,1,1,2}$	0	1	1	2	✓	✓	✓
	$c_{0,0,1,2}$	0	0	1	2	✓	✓	✓
	$c_{0,2,0,2}$	0	2	0	2	✓	✓	
	$c_{0,1,0,2}$	0	1	0	2	✓	✓	
	$c_{0,0,0,2}$	0	0	0	2	✓	✓	✓
$cd_{2,1}$	$d_{2,2,2,1}$	2	2	2	1			
	$d_{2,1,2,1}$	2	1	2	1			
	$d_{2,0,2,1}$	2	0	2	1			
	$d_{2,2,1,1}$	2	2	1	1			
	$d_{2,1,1,1}$	2	1	1	1			
	$d_{2,0,1,1}$	2	0	1	1			
	$d_{2,2,0,1}$	2	2	0	1			
	$d_{2,1,0,1}$	2	1	0	1			
	$d_{2,0,0,1}$	2	0	0	1			
$cn_{1,1}$	$n_{1,2,2,1}$	1	2	2	1	✓		
	$n_{1,1,2,1}$	1	1	2	1	✓		
	$n_{1,0,2,1}$	1	0	2	1	✓		
	$n_{1,2,1,1}$	1	2	1	1	✓		
	$n_{1,1,1,1}$	1	1	1	1	✓	✓	✓
	$n_{1,0,1,1}$	1	0	1	1	✓		
	$n_{1,2,0,1}$	1	2	0	1	✓		
	$n_{1,1,0,1}$	1	1	0	1	✓		
	$n_{1,0,0,1}$	1	0	0	1	✓		
$cc_{0,1}$	$c_{0,2,2,1}$	0	2	2	1	✓		
	$c_{0,1,2,1}$	0	1	2	1	✓		
	$c_{0,0,2,1}$	0	0	2	1	✓		
	$c_{0,2,1,1}$	0	2	1	1	✓		
	$c_{0,1,1,1}$	0	1	1	1	✓	✓	✓
	$c_{0,0,1,1}$	0	0	1	1	✓	✓	✓
	$c_{0,2,0,1}$	0	2	0	1	✓		
	$c_{0,1,0,1}$	0	1	0	1	✓	✓	
	$c_{0,0,0,1}$	0	0	0	1	✓	✓	✓
$cd_{2,0}$	$d_{2,2,2,0}$	2	2	2	0			

Continued on next page

Table 10 – continued from previous page

Combined type	Type	D^{00}	D^{10}	D^{01}	D^{11}	LiM	PM	IAM
	$d_{2,1,2,0}$	2	1	2	0			
	$d_{2,0,2,0}$	2	0	2	0			
	$d_{2,2,1,0}$	2	2	1	0			
	$d_{2,1,1,0}$	2	1	1	0			
	$d_{2,0,1,0}$	2	0	1	0			
	$d_{2,2,0,0}$	2	2	0	0			
	$d_{2,1,0,0}$	2	1	0	0			
	$d_{2,0,0,0}$	2	0	0	0			
$cd_{1,0}$	$d_{1,2,2,0}$	1	2	2	0			
	$d_{1,1,2,0}$	1	1	2	0			
	$d_{1,0,2,0}$	1	0	2	0			
	$d_{1,2,1,0}$	1	2	1	0			
	$d_{1,1,1,0}$	1	1	1	0			
	$d_{1,0,1,0}$	1	0	1	0			
	$d_{1,2,0,0}$	1	2	0	0			
	$d_{1,1,0,0}$	1	1	0	0			
	$d_{1,0,0,0}$	1	0	0	0			
$cn_{0,0}$	$n_{0,2,2,0}$	0	2	2	0	✓		
	$n_{0,1,2,0}$	0	1	2	0	✓		
	$n_{0,0,2,0}$	0	0	2	0	✓		
	$n_{0,2,1,0}$	0	2	1	0	✓		
	$n_{0,1,1,0}$	0	1	1	0	✓		
	$n_{0,0,1,0}$	0	0	1	0	✓		
	$n_{0,2,0,0}$	0	2	0	0	✓		
	$n_{0,1,0,0}$	0	1	0	0	✓		
	$n_{0,0,0,0}$	0	0	0	0	✓	✓	✓

B Proofs

B.1 Proof of Theorem 1

First note that $\sum_{k,l} P(T = cc_{k,l}) = \sum_{k \leq l} P(T = cc_{k,l}) = 1$ because of LiM. Then, consider the first part of the numerator of $\beta \equiv \frac{E(Y|Z_1=Z_2=...=Z_K=1)-E(Y|Z_1=Z_2=...=Z_K=0)}{E(D|Z_1=Z_2=...=Z_K=1)-E(D|Z_1=Z_2=...=Z_K=0)}$:

$$\begin{aligned}
E(Y|Z_1 = Z_2 = ... = Z_K = 1) &= E(Y|\widetilde{Z} = 1) \\
&= \sum_{k,l} E(Y|\widetilde{Z} = 1, T = cc_{k,l}) P(T = cc_{k,l}|\widetilde{Z} = 1) \\
&= \sum_{k \leq l} E(Y|\widetilde{Z} = 1, T = cc_{k,l}) P(T = cc_{k,l}|\widetilde{Z} = 1) \\
&= \sum_{k \leq l} E(Y^l|T = cc_{k,l}) P(T = cc_{k,l}),
\end{aligned}$$

where the third equality follows from LiM and the last equality follows from the exclusion and unconfoundedness/independence assumptions.

Similarly, consider the other components of

$$E(Y|Z_1 = Z_2 = ... = Z_K = 0) = E(Y|\widetilde{Z} = 0) = \sum_{k \leq l} E(Y^k|T = c_{k,l}) P(T = cc_{k,l}),$$

and

$$E(D|Z_1 = Z_2 = ... = Z_K = 1) = E(D|\widetilde{Z} = 1) = \sum_{k \leq l} l \cdot P(T = cc_{k,l}),$$

and

$$E(D|Z_1 = Z_2 = ... = Z_K = 0) = E(D|\widetilde{Z} = 0) = \sum_{k \leq l} k \cdot P(T = cc_{k,l}).$$

Combining the above results:

$$\begin{aligned}
&\frac{E(Y|\widetilde{Z} = 1) - E(Y|\widetilde{Z} = 0)}{E(D|\widetilde{Z} = 1) - E(D|\widetilde{Z} = 0)} \\
&= \frac{\sum_{k \leq l} E(Y^l|T = cc_{k,l}) P(T = cc_{k,l}) - \sum_{k \leq l} E(Y^k|T = cc_{k,l}) P(T = cc_{k,l})}{\sum_{k \leq l} l \cdot P(T = cc_{k,l}) - \sum_{k \leq l} k \cdot P(T = cc_{k,l})} \\
&= \frac{\sum_{k \leq l} E(Y^l - Y^k|T = cc_{k,l}) P(T = cc_{k,l})}{\sum_{k \leq l} (l - k) \cdot P(T = cc_{k,l})} \\
&= \frac{\sum_{k \leq l} E(Y^l - Y^k|T = cc_{k,l}) P(T = cc_{k,l})}{\sum_{k \leq l} (l - k) \cdot P(T = cc_{k,l})} \\
&= \frac{\sum_{k \leq l} E(Y^l - Y^k|T = cc_{k,l}) P(T = cc_{k,l})}{\sum_{k \leq l} (l - k) \cdot P(T = cc_{k,l})} \\
&= \frac{\sum_{k < l} E(Y^l - Y^k|T = cc_{k,l}) P(T = cc_{k,l})}{\sum_{k < l} (l - k) \cdot P(T = cc_{k,l})} + \frac{\sum_{k = l} E(Y^l - Y^k|T = cc_{k,l}) P(T = cc_{k,l})}{\sum_{k = l} (l - k) \cdot P(T = cc_{k,l})} \\
&= \frac{\sum_{k < l} E(Y^l - Y^k|T = cc_{k,l}) P(T = cc_{k,l})}{\sum_{k < l} (l - k) \cdot P(T = cc_{k,l})}
\end{aligned}$$

$$= \sum_{k<l} \frac{P(T = cc_{k,l})}{\sum_{k<l}(l-k) \cdot P(T = cc_{k,l})} E(Y^l - Y^k | T = cc_{k,l}).$$

B.2 Proof of Proposition 1

Suppose that the treatment D is discrete with bounded support. Denote with M the number of elements in the rectangular instrument support $\mathcal{Z}$ ordered such that $l < m$ implies $E(D|Z = l) < E(D|Z = m)$. Label the ordered elements as z_1, z_2, ..., z_M. Theorem 2 of Angrist and Imbens (1995) establishes that TSLS combined with Assumptions 1 to 3 estimates

$$\beta_{TSLS} \equiv \sum_{m=1}^{M} \mu_m \cdot \beta_{m,m-1},$$

where

$$\mu_m = (E(D|Z = z_m) - E(D|Z = z_{m-1})) \cdot \frac{\sum_{l=m}^{M} P(Z = z_l)(E(D|Z = z_l) - E(D))}{\sum_{l=0}^{M} P(Z = z_l)E(D|Z = z_l)(E(D|Z = z_l) - E(D))}$$

and

$$\beta_{m,m-1} = \frac{E(Y|Z = z_m) - E(Y|Z = z_{m-1})}{E(D|Z = z_m) - E(D|Z = z_{m-1})}.$$

Using this as a starting point, I now show that this can be rewritten to obtain an interpretation of a weighted average of causal responses for different response types:

$$\begin{aligned}
&\beta_{TSLS} \\
&= \sum_{m=1}^{M} (E(D|Z = z_m) - E(D|Z = z_{m-1})) \cdot \frac{\sum_{l=m}^{M} P(Z = z_l)(E(D|Z = z_l) - E(D))}{\sum_{l=0}^{M} P(Z = z_l)E(D|Z = z_l)(E(D|Z = z_l) - E(D))} \\
&\quad \cdot \frac{E(Y|Z = z_m) - E(Y|Z = z_{m-1})}{E(D|Z = z_m) - E(D|Z = z_{m-1})} \\
&= \sum_{m=1}^{M} \frac{\sum_{l=m}^{M} P(Z = z_l)(E(D|Z = z_l) - E(D))}{\sum_{l=0}^{M} P(Z = z_l)E(D|Z = z_l)(E(D|Z = z_l) - E(D))} \cdot E(Y|Z = z_m) - E(Y|Z = z_{m-1}) \\
&= \sum_{m=1}^{M} \frac{\sum_{l=m}^{M} P(Z = z_l)(E(D|Z = z_l) - E(D))}{\sum_{l=0}^{M} P(Z = z_l)E(D|Z = z_l)(E(D|Z = z_l) - E(D))} \cdot E(Y^{D^{z_m}} - Y^{D^{z_{m-1}}}) \\
&\equiv \sum_{m=1}^{M} \omega_m \cdot E(Y^{D^{z_m}} - Y^{D^{z_{m-1}}}),
\end{aligned}$$

where the weights are

$$\omega_m = \frac{\sum_{l=m}^{M} P(Z = z_l)(E(D|Z = z_l) - E(D))}{\sum_{l=0}^{M} P(Z = z_l)E(D|Z = z_l)(E(D|Z = z_l) - E(D))}.$$

Denote $\mathcal{T}$ the set of all response types t, and $\sum_{t \in \mathcal{T}} P(T = t) = 1$. Further denote $I(\cdot)$ the indicator function, which equals one if its argument is true and zero otherwise. Then, it can be

shown that TSLS preserves the interpretation of a weighted average of causal responses $Y^a - Y^b$ where $a > b$:

$$
\begin{aligned}
\beta_{TSLS} &= \sum_{m=1}^{M} \omega_m \cdot E(Y^{D^{z_m}} - Y^{D^{z_{m-1}}}) \\
&= \sum_{m=1}^{M} \omega_m \left(\sum_{t \in \mathcal{T}} P(T=t) \cdot E(Y^{D^{z_m}} - Y^{D^{z_{m-1}}} | T=t) \right) \\
&= \sum_{t \in \mathcal{T}} \left(P(T=t) \sum_{m=1}^{M} \omega_m \cdot E(Y^{D^{z_m}} - Y^{D^{z_{m-1}}} | T=t) \right) \\
&= \sum_{t \in \mathcal{T}} \Big(P(T=t) \sum_{m=1}^{M} \Big\{ I(D^{z_m} > D^{z_{m-1}}) \cdot \omega_m \cdot E(Y^{D^{z_m}} - Y^{D^{z_{m-1}}} | T=t) \\
&\qquad\qquad - I(D^{z_m} < D^{z_{m-1}}) \cdot \omega_m \cdot E(Y^{D^{z_{m-1}}} - Y^{D^{z_m}} | T=t) \Big\} \Big) \\
&\equiv \sum_{t \in \mathcal{T}} P(T=t) \sum_{m=1}^{M} \iota_{m,m-1} \cdot \omega_m \cdot E(Y^{D^{z_m}} - Y^{D^{z_{m-1}}} | T=t),
\end{aligned}
$$

where

$$
\iota_{m,m-1} \equiv I(D^{z_m} \geq D^{z_{m-1}}) - I(D^{z_m} \leq D^{z_{m-1}}) = \begin{cases} -1 & \text{if } D^{z_m} < D^{z_{m-1}} \\ 1 & \text{if } D^{z_m} > D^{z_{m-1}} \\ 0 & \text{if } D^{z_m} = D^{z_{m-1}} \end{cases},
$$

and

$$
\omega_m = \frac{\sum_{l=m}^{M} P(Z=z_l)(E(D|Z=z_l) - E(D))}{\sum_{l=0}^{M} P(Z=z_l)E(D|Z=z_l)(E(D|Z=z_l) - E(D))}.
$$

The fourth equality holds since $Y^{D^{z_m}} - Y^{D^{z_{m-1}}} = 0$ when $D^{z_m} = D^{z_{m-1}}$. Note that the numerator of ω_m can be re-written as follows:

$$
\begin{aligned}
&\sum_{l=m}^{M} P(Z=z_l)(E(D|Z=z_l) - E(D)) \\
&= \sum_{l=m}^{M} (P(Z=z_l)E(D|Z=z_l)) - \sum_{l=m}^{M} (P(Z=z_l)E(D)) \\
&= \sum_{l=m}^{M} (P(Z=z_l)E(D|Z=z_l)) - E(D) \sum_{l=m}^{M} P(Z=z_l) \\
&= \sum_{l=m}^{M} (P(Z=z_l)E(D|Z=z_l)) - P(Z \geq k)E(D) \\
&= \sum_{l=m}^{M} (P(Z=z_l)E(D|Z=z_l)) - P(Z \geq z_k)(P(Z < z_k)E(D|Z < z_k) + P(Z \geq z_k)E(D|Z \geq z_k)) \\
&= (1 - P(Z \geq z_k))P(Z \geq z_k)E(D|Z \geq z_k) - P(Z \geq z_k)P(Z < z_k)E(D|Z < z_k) \\
&= (1 - P(Z \geq z_k))P(Z \geq z_k)E(D|Z \geq z_k) - P(Z \geq z_k)(1 - P(Z \geq z_k))E(D|Z < z_k)
\end{aligned}
$$

$$= (1 - P(Z \geq z_k))P(Z \geq z_k) \cdot \{E(D|Z \geq z_k) - E(D|Z < z_k)\}.$$

B.3 Proof of alternative formulation of Theorem 1

Write Y as follows:

$$\begin{aligned}
Y &= I(Z_1 = Z_2 = ... = Z_K = 1) \cdot Y^{D^{1...1...1}} + I(Z_1 = Z_2 = ... = Z_K = 0) \cdot Y^{D^{0...0...0}} \\
&\quad + I(Z_1 = q, ..., Z_k = r, ..., Z_K = s) \cdot Y^{D^{q...r...s}} \\
&= \left(I(Z_1 = Z_2 = ... = Z_K = 1) \cdot \sum_{j=0}^{J} Y^j \cdot I(D^{1...1...1} \geq j) \right) \\
&\quad + \left(I(Z_1 = Z_2 = ... = Z_K = 0) \cdot \sum_{j=0}^{J} Y^j \cdot I(D^{0...0...0} \geq j) \right) \\
&\quad + \left(I(Z_1 = q, ..., Z_k = r, ..., Z_K = s) \cdot \sum_{j=0}^{J} Y^j \cdot I(D^{q...r...s} \geq j) \right),
\end{aligned}$$

$\forall q, r, s$ such that $q \neq r \neq s$.

First, consider the numerator in $\beta_{\text{CC-ACR}} \equiv \frac{E(Y|Z_1=Z_2=...=Z_K=1)-E(Y|Z_1=Z_2=...=Z_K=0)}{E(D|Z_1=Z_2=...=Z_K=1)-E(D|Z_1=Z_2=...=Z_K=0)}$:

$$\begin{aligned}
&E(Y|Z_1 = Z_2 = ... = Z_K = 1) - E(Y|Z_1 = Z_2 = ... = Z_K = 0) \\
&= E\left(\sum_{j=0}^{J} Y^j \cdot I(D^{1...1...1} \geq j) | Z_1 = Z_2 = ... = Z_K = 1 \right) \\
&\quad - E\left(\sum_{j=0}^{J} Y^j \cdot I(D^{0...0...0} \geq j) | Z_1 = Z_2 = ... = Z_K = 0 \right) \\
&= E\left(\sum_{j=0}^{J} Y^j \cdot (I(D^{1...1...1} \geq j) - I(D^{0...0...0} \geq j)) \right) \\
&= E\left(\sum_{j=0}^{J} Y^j \cdot (I(D^{1...1...1} \geq j) - I(D^{1...1...1} \geq j+1) - I(D^{0...0...0} \geq j) - I(D^{0...0...0} \geq j+1)) \right) \\
&= E\Bigg(Y_0 \cdot (I(D^{1...1...1} \geq 0) - I(D^{1...1...1} \geq 1) - I(D^{0...0...0} \geq 0) - I(D^{0...0...0} \geq 1)) \\
&\quad + \sum_{j=1}^{J} Y^j \cdot (I(D^{1...1...1} \geq j) - I(D^{1...1...1} \geq j+1) - I(D^{0...0...0} \geq j) - I(D^{0...0...0} \geq j+1)) \Bigg) \\
&= E\left(Y_0 \cdot (I(D^{1...1...1} \geq 0) - I(D^{0...0...0} \geq 0)) + \sum_{j=1}^{J} (Y^j - Y^{j-1}) \cdot (I(D^{1...1...1} \geq j) - I(D^{0...0...0} \geq j)) \right) \\
&= E\left(\sum_{j=1}^{J} (Y^j - Y^{j-1}) \cdot (I(D^{1...1...1} \geq j) - I(D^{0...0...0} \geq j)) \right).
\end{aligned}$$

$I(D^{1...1...1} \geq j) - I(D^{0...0...0} \geq j)$ equals zero or one since $I(D^{1...1...1} \geq j) \geq I(D^{0...0...0} \geq j)$.

Subsequently:

$$\sum_{j=1}^{J} E(Y^j - Y^{j-1} | I(D^{1...1...1} \geq j) - I(D^{0...0...0} \geq j) = 1) \cdot P(I(D^{1...1...1} \geq j) - I(D^{0...0...0} \geq j) = 1)$$
$$= \sum_{j=1}^{J} E(Y^j - Y^{j-1} | D^{1...1...1} \geq j > D^{0...0...0}) \cdot P(D^{1...1...1} \geq j > D^{0...0...0}).$$

Now, write D as follows:

$$\begin{aligned} D &= I(Z_1 = Z_2 = ... = Z_K = 1) \cdot D^{1...1...1} + I(Z_1 = Z_2 = ... = Z_K = 0) \cdot D^{0...0...0} \\ &\quad + I(Z_1 = q, ..., Z_k = r, ..., Z_K = s) \cdot D^{q...r...s} \\ &= \left(I(Z_1 = Z_2 = ... = Z_K = 1) \cdot \sum_{j=0}^{J} j \cdot I(D^{1...1...1} \geq j) \right) \\ &\quad + \left(I(Z_1 = Z_2 = ... = Z_K = 0) \cdot \sum_{j=0}^{J} j \cdot I(D^{0...0...0} \geq j) \right) \\ &\quad + \left(I(Z_1 = q, ..., Z_k = r, ..., Z_K = s) \cdot \sum_{j=0}^{J} j \cdot I(D^{q...r...s} \geq j) \right), \end{aligned}$$

$\forall q, r, s$ such that $q \neq r \neq s$.

Then, consider the denominator in $\beta_{\text{CC-ACR}} \equiv \frac{E(Y|Z_1=Z_2=...=Z_K=1)-E(Y|Z_1=Z_2=...=Z_K=0)}{E(D|Z_1=Z_2=...=Z_K=1)-E(D|Z_1=Z_2=...=Z_K=0)}$:

$$\begin{aligned} &E(D|Z_1 = Z_2 = ... = Z_K = 1) - E(D|Z_1 = Z_2 = ... = Z_K = 0) \\ &= E\left(\sum_{j=0}^{J} j \cdot I(D^{1...1...1} \geq j) | Z_1 = Z_2 = ... = Z_K = 1 \right) \\ &\quad - E\left(\sum_{j=0}^{J} j \cdot I(D^{0...0...0} \geq j) | Z_1 = Z_2 = ... = Z_K = 0 \right) \\ &= E\left(\sum_{j=0}^{J} j \cdot (I(D^{1...1...1} = j) - I(D^{0...0...0} = j)) \right) \\ &= E\left(\sum_{j=0}^{J} j \cdot (I(D^{1...1...1} \geq j) - I(D^{1...1...1} \geq j+1) - I(D^{0...0...0} \geq j) - I(D^{0...0...0} \geq j+1)) \right) \\ &= E\left(\sum_{j=1}^{J} I(D^{1...1...1} \geq j) - I(D^{0...0...0} \geq j) \right) \\ &= \sum_{j=1}^{J} P(D^{1...1...1} \geq j > D^{0...0...0}). \end{aligned}$$

It is required that $P(D^{1...1...1} \geq j > D^{0...0...0}) > 0$ for some j which imposes a relevance assumption on the instrument. Moreover, $P(D^{1...1...1} \geq l > D^{0...0...0}) = \sum_{l>k} P(T = c_{l,k})$.

Then:

$$\frac{E(Y|Z_1 = Z_2 = ... = Z_K = 1) - E(Y|Z_1 = Z_2 = ... = Z_K = 0)}{E(D|Z_1 = Z_2 = ... = Z_K = 1) - E(D|Z_1 = Z_2 = ... = Z_K = 0)}$$
$$= \sum_{j=1}^{J} \frac{P(D^{1...1...1} \geq j > D^{0...0...0})}{\sum_{i=1}^{J} P(D^{1...1...1} \geq i > D^{0...0...0})} E(Y^j - Y^{j-1}|D^{1...1...1} \geq j > D^{0...0...0})$$
$$= \sum_{j=1}^{J} \frac{P(T = c_{l,k})}{\sum_{i=1}^{J} \sum_{l>i} P(T = c_{l,i})} E(Y^j - Y^{j-1}|D^{1...1...1} \geq j > D^{0...0...0}).$$

B.4 Proof of alternative representation of the TSLS estimand

Write Theorem 2 of Angrist and Imbens (1995) as follows:

$$\beta_{TSLS} \equiv \sum_{m=1}^{M} \delta_{m,m-1} \cdot \omega_m \cdot \beta_{m,m-1}, \tag{11}$$

where

$$\delta_{m,m-1} = E(D|Z = z_m) - E(D|Z = z_{m-1}),$$

and

$$\omega_m = \frac{\sum_{l=m}^{M} P(Z = z_l)(E(D|Z = z_l) - E(D))}{\sum_{l=0}^{M} P(Z = z_l)E(D|Z = z_l)(E(D|Z = z_l) - E(D))},$$

and

$$\beta_{m,m-1} = \frac{E(Y|Z = z_m) - E(Y|Z = z_{m-1})}{E(D|Z = z_m) - E(D|Z = z_{m-1})}.$$

It holds that

$$E(Y|Z = z_m) - E(Y|Z = z_{m-1}) = \sum_{j=1}^{J} P(D^{z_m} \geq j > D^{z_{m-1}}) \cdot E(Y^j - Y^{j-1}|D^{z_m} \geq j > D^{z_{m-1}})$$
$$+ \sum_{j=1}^{J} P(D^{z_m} < j \leq D^{z_{m-1}}) \cdot E(Y^j - Y^{j-1}|D^{z_m} < j \leq D^{z_{m-1}}).$$

This can be seen as follows:

$$E(Y|Z = z_m) - E(Y|Z = z_{m-1})$$
$$= E\left(\sum_{j=0}^{J} Y^j \cdot I(D^{z_m} \geq j)|Z = z_m\right) - E\left(\sum_{j=0}^{J} Y^j \cdot I(D^{z_{m-1}} \geq j)|Z = z_{m-1}\right)$$
$$= E\left(\sum_{j=0}^{J} Y^j \cdot (I(D^{z_m} \geq j) - I(D^{z_{m-1}} \geq j))\right)$$
$$= E\left(\sum_{j=0}^{J} Y^j \cdot (I(D^{z_m} \geq j) - I(D^{z_m} \geq j+1) - I(D^{z_{m-1}} \geq j) - I(D^{z_{m-1}} \geq j+1))\right)$$

$$
\begin{aligned}
&= E\Bigg(Y^{j-1} \cdot (I(D^{z_m} \geq 0) - I(D^{z_m} \geq 1) - I(D^{z_{m-1}} \geq j-1) - I(D^{z_{m-1}} \geq 1)) \\
&\quad + \sum_{j=1}^{J} Y^j \cdot (I(D^{z_m} \geq j) - I(D^{z_m} \geq j+1) - I(D^{z_{m-1}} \geq j) - I(D^{z_{m-1}} \geq j+1))\Bigg) \\
&= E\Bigg(Y^{j-1} \cdot (I(D^{z_m} \geq j-1) - I(D^{z_{m-1}} \geq j-1)) \\
&\quad + \sum_{j=1}^{J} (Y^j - Y^{j-1}) \cdot (I(D^{z_m} \geq j) - I(D^{z_{m-1}} \geq j))\Bigg) \\
&= E(\sum_{j=1}^{J} (Y^j - Y^{j-1}) \cdot (I(D^{z_m} \geq j) - I(D^{z_{m-1}} \geq j))) \\
&= \sum_{j=1}^{J} E((Y^j - Y^{j-1}) \cdot (I(D^{z_m} \geq j) - I(D^{z_{m-1}} \geq j))) \\
&= \sum_{j=1}^{J} E\left(E((Y^j - Y^{j-1}) \cdot (I(D^{z_m} \geq j) - I(D^{z_{m-1}} \geq j)) | I(D^{z_m} \geq j) - I(D^{z_{m-1}} \geq j))\right) \\
&= \sum_{j=1}^{J} \Big\{1 \cdot P(I(D^{z_m} \geq j) - I(D^{z_{m-1}} \geq j) = 1) \\
&\qquad \cdot E((Y^j - Y^{j-1}) \cdot (I(D^{z_m} \geq j) - I(D^{z_{m-1}} \geq j)) | I(D^{z_m} \geq j) - I(D^{z_{m-1}} \geq j) = 1) \\
&\quad + 0 \cdot P(I(D^{z_m} \geq j) - I(D^{z_{m-1}} \geq j) = 0) \\
&\qquad \cdot E((Y^j - Y^{j-1}) \cdot (I(D^{z_m} \geq j) - I(D^{z_{m-1}} \geq j)) | I(D^{z_m} \geq j) - I(D^{z_{m-1}} \geq j) = 0) \\
&\quad - 1 \cdot P(I(D^{z_m} \geq j) - I(D^{z_{m-1}} \geq j) = -1) \\
&\qquad \cdot E((Y^j - Y^{j-1}) \cdot (I(D^{z_m} \geq j) - I(D^{z_{m-1}} \geq j)) | I(D^{z_m} \geq j) - I(D^{z_{m-1}} \geq j) = -1)\Big\} \\
&= 1 \cdot \sum_{j=1}^{J} P(D^{z_m} \geq j > D^{z_{m-1}}) \cdot E((Y^j - Y^{j-1}) \cdot 1 | D^{z_m} \geq j > D^{z_{m-1}}) \\
&\quad - 1 \cdot \sum_{j=1}^{J} P(D^{z_m} < j \leq D^{z_{m-1}}) \cdot E((Y^j - Y^{j-1}) \cdot (-1) | D^{z_m} < j \leq D^{z_{m-1}}) \\
&= \sum_{j=1}^{J} P(D^{z_m} \geq j > D^{z_{m-1}}) \cdot E(Y^j - Y^{j-1} | D^{z_m} \geq j > D^{z_{m-1}}) \\
&\quad - \sum_{j=1}^{J} P(D^{z_{m-1}} \geq j > D^{z_m}) \cdot E(Y^{j-1} - Y^j | D^{z_{m-1}} \geq j > D^{z_m}).
\end{aligned}
$$

Now, it rests to show that

$$
E(D|Z = z_m) - E(D|Z = z_{m-1}) = \sum_{i=1}^{J} P(D^{z_m} \geq i > D^{z_{m-1}}).
$$

This can be shown as follows:

$$
D = Z \cdot D^{z_m} + (1 - Z) D^{z_{m-1}} = \left(Z \cdot \sum_{j=0}^{J} j \cdot I(D^{z_m} \geq j)\right) + \left((1 - Z) \cdot \sum_{j=0}^{J} j \cdot I(D^{z_{m-1}} \geq j)\right).
$$

Then:

$$
\begin{aligned}
&E(D|Z=z_m) - E(D|Z=z_{m-1}) \\
&= E\left(\sum_{j=0}^{J} j \cdot I(D^{z_m} \geq j)|Z=z_m\right) - E\left(\sum_{j=0}^{J} j \cdot I(D^{z_{m-1}} \geq j)|Z=z_{m-1}\right) \\
&= E\left(\sum_{j=0}^{J} j(I(D^{z_m}=j) - I(D^{z_{m-1}}=j))\right) \\
&= E\left(\sum_{j=0}^{J} j \cdot (I(D^{z_m} \geq j) - I(D^{z_m} \geq j+1) - I(D^{z_{m-1}} \geq j) - I(D^{z_{m-1}} \geq j+1))\right) \\
&= E\left(\sum_{j=1}^{J} I(D^{z_m} \geq j) - I(D^{z_{m-1}} \geq j)\right) \\
&= \sum_{j=1}^{J} P(D^{z_m} \geq j > D^{z_{m-1}}).
\end{aligned}
$$

It is required that $P(D^{z_m} \geq j > D^{z_{m-1}}) > 0$ for some j which imposes a relevance assumption on the instrument.

Plugging the above results into $\beta_{m,m-1}$, we get:

$$
\begin{aligned}
\beta_{m,m-1} &= \frac{E(Y|Z=z_m) - E(Y|Z=z_{m-1})}{E(D|Z=z_m) - E(D|Z=z_{m-1})} \\
&= \sum_{j=1}^{J} \frac{P(D^{z_m} \geq j > D^{z_{m-1}})}{\sum_{i=1}^{J} P(D^{z_m} \geq i > D^{z_{m-1}})} \cdot E(Y^j - Y^{j-1}|D^{z_m} \geq j > D^{z_{m-1}}) \\
&\quad - \sum_{j=1}^{J} \frac{P(D^{z_{m-1}} \geq j > D^{z_m})}{\sum_{i=1}^{J} P(D^{z_m} \geq i > D^{z_{m-1}})} \cdot E(Y^{j-1} - Y^j|D^{z_{m-1}} \geq j > D^{z_m}).
\end{aligned}
$$

Then, Equation (11) without imposing any monotonicity can be re-written to:

$$
\begin{aligned}
\beta_{\text{TSLS,PM}} \equiv \sum_{m=1}^{M} \Big\{ & I(D^{z_m} > D^{z_{m-1}}) \cdot \delta_{m,m-1} \cdot \omega_m \cdot \beta^c_{m,m-1} \\
& - I(D^{z_m} < D^{z_{m-1}}) \cdot \delta_{m,m-1} \cdot \omega_m \cdot \beta^d_{m,m-1} \Big\},
\end{aligned}
$$

where

$$
\delta_{m,m-1} = E(D|Z=z_m) - E(D|Z=z_{m-1}),
$$

and

$$
\omega_m = \frac{\sum_{l=m}^{M} P(Z=z_l)(E(D|Z=z_l) - E(D))}{\sum_{l=0}^{M} P(Z=z_l)E(D|Z=z_l)(E(D|Z=z_l) - E(D))},
$$

and

$$
\beta^c_{m,m-1} = \sum_{j=1}^{J} \frac{P(D^{z_m} \geq j > D^{z_{m-1}})}{\sum_{i=1}^{J} P(D^{z_m} \geq i > D^{z_{m-1}})} E(Y^j - Y^{j-1}|D^{z_m} \geq j > D^{z_{m-1}}),
$$

and

$$
\beta^d_{m,m-1} = \sum_{j=1}^{J} \frac{P(D^{z_{m-1}} \geq j > D^{z_m})}{\sum_{i=1}^{J} P(D^{z_m} \geq i > D^{z_{m-1}})} \cdot E(Y^j - Y^{j-1}|D^{z_{m-1}} \geq j > D^{z_m}),
$$

where the superscripts c and d denote whether β gives the LATE for those who respond as compliers or defiers for a change from $m-1$ to m respectively.

The weights ω_m are equivalent to the presentation of the previous section. The weights $\delta_{m,m-1} \cdot \omega_m$ sum to one ($\sum_{m=1}^{M} \delta_{m,m-1} \cdot \omega_m = 1$), and are non-negative ($\delta_{m,m-1} \cdot \omega_m > 0$ for all m). The weights are proportional to the impact that the instrument with k used in constructing $\beta_{k,k-1}$ has on the treatment level. Similar to the previous section, more weight is given to $E(Y^j - Y^{j-1}|D^{z_m} \geq j > D^{z_{m-1}})$ and $E(Y^j - Y^{j-1}|D^{z_{m-1}} \geq j > D^{z_m})$ if it lies in the center of the instrument distribution.

B.5 Proof for a continuous treatment

Combining the arguments in the present study with those of Angrist, Graddy, and Imbens (2000), the following can be shown:

$$
\begin{aligned}
&E(Y|Z_1 = Z_2 = ... = Z_K = 1) - E(Y|Z_1 = Z_2 = ... = Z_K = 0) \\
&= E(Y^{D^{1...1...1}} - Y^{D^{0...0...0}}) \\
&= E\left(\int_0^{D^{1...1...1}} \frac{\partial Y^t}{\partial t} dt - \int_{D^{0...0...0}}^{\infty} \frac{\partial Y^t}{\partial t} dt\right) \\
&= E\left(\int_{D^{0...0...0}}^{D^{1...1...1}} \frac{\partial Y^t}{\partial t} dt\right) \\
&= E\left(\int_0^{\infty} I\{D^{1...1...1} \geq t > D^{0...0...0}\} \frac{\partial Y^t}{\partial t} dt\right) \\
&= \int_0^{\infty} E\left(I\{D^{1...1...1} \geq t > D^{0...0...0}\} \frac{\partial Y^t}{\partial t}\right) dt \\
&= \int_0^{\infty} P(D^{1...1...1} \geq t > D^{0...0...0}) \cdot \frac{\partial E(Y^t|D^{1...1...1} \geq t > D^{0...0...0})}{\partial t} dt.
\end{aligned}
$$

The independence assumption and the fundamental theorem of calculus ($f(x) = \int_0^x f'(t)dt = \int_0^x \frac{\partial f(t)}{\partial t} dt$) were used in lines two and three, respectively. Similarly, it can be shown that

$$
\begin{aligned}
&E(D|Z_1 = Z_2 = ... = Z_K = 1) - E(D|Z_1 = Z_2 = ... = Z_K = 0) \\
&= \int_0^{\infty} P(D^{1...1...1} \geq j > D^{0...0...0}) dj.
\end{aligned}
$$

Then:

$$
\begin{aligned}
\beta_{\text{CC-ACR}} &= \frac{E(Y|Z_1 = Z_2 = ... = Z_K = 1) - E(Y|Z_1 = Z_2 = ... = Z_K = 0)}{E(D|Z_1 = Z_2 = ... = Z_K = 1) - E(D|Z_1 = Z_2 = ... = Z_K = 0)} \\
&= \frac{\int_0^{\infty} P(D^{1...1...1} \geq t > D^{0...0...0}) \cdot \frac{\partial E(Y^t|D^{1...1...1} \geq t > D^{0...0...0})}{\partial t} dt}{\int_0^{\infty} P(D^{1...1...1} \geq j > D^{0...0...0}) dj}.
\end{aligned}
$$

C LiM test

C.1 Inequalities for detecting local violations of LiM

This section shows how the inequalities for the LiM test can be derived. As the LiM assumption provides the condition that the CDFs do not cross, which is equivalent to the condition of having positive weights at every point of the distribution of D, the following $J+1$ inequalities have to hold under LiM (which can be derived from Equation (7)):

$$P(D < j | Z_1 = Z_2 = ... = Z_K = 0) - P(D < j | Z_1 = Z_2 = ... = Z_K = 1) \geq 0, \tag{12}$$

for all $j \in \{0, 1, ..., J\}$.

The inequalities in Condition (12) translate to learning the sign of the causal effect on the treatment variable of the sole instrument, $\widetilde{Z} = Z_1 = Z_2 = ... = Z_K$, in the subsample of observations at the outer support of the instrument values:

$$P(D < j | \widetilde{Z} = 0) - P(D < j | \widetilde{Z} = 1) \geq 0 \text{ for all } j \in \{0, 1, ..., J\}.$$

Rewrite the previous equation to the following expression:

$$E(I(D < j) | \widetilde{Z} = 0) - E(I(D < j) | \widetilde{Z} = 1) \geq 0 \text{ for all } j \in \{0, 1, ..., J\}.$$

Then, the following inequality must be satisfied at any point x in the covariate space:

$$E(I(D < j) | \widetilde{Z} = 0, X = x) - E(I(D < j) | \widetilde{Z} = 1, X = x) \geq 0 \text{ for all } j \in \{0, 1, ..., J\}.$$

C.2 Test procedure

The procedure by Farbmacher, Guber, and Klaassen (2022) can be followed for estimating $\tau_j(x)$ of Equation (9) and is described here. The average treatment effect given by this equation gives an insight into the magnitude of possible violations. Two additional assumptions are required to establish causality: (1) $(Q_J^1, Q_J^0) \perp \widetilde{Z} | X$, and (2) $\epsilon < P(\widetilde{Z} = 1 | X = x) < 1 - \epsilon$ for some $\epsilon > 0$. Then, the average treatment effect can be estimated using augmented inverse-propensity weighting based on Robins, Rotnitzky, and Zhao (1994):

$$\begin{aligned} \hat{\Gamma}_{j,i} \equiv & \hat{\tau}_j^{(-i)}(X_i) \\ & + \frac{\widetilde{Z}_i - \hat{e}^{(-i)}(X_i)}{\hat{e}^{(-i)}(X_i)\left(1 - \hat{e}^{(-i)}(X_i)\right)} \times \left(Q_{j,i} - \hat{\mu}_j^{(-i)}(X_i) - (\widetilde{Z}_i - \hat{e}^{(-i)}(X_i))\hat{\tau}_j^{(-i)}(X_i)\right). \end{aligned} \tag{13}$$

$\hat{\tau}_j(X_i)$, $\hat{e}(X_i)$, and $\hat{\mu}_j(X_i)$ are estimates of $\tau_j(x)$, $e(x) = P(\widetilde{Z}_i = 1 | X_i = x)$, and $\mu_j(x) = E(Q_{j,i} | X_i = x)$, respectively. The superscript $(-i)$ denotes out-of-bag estimates. This means that estimates were obtained without the ith observation (e.g., D_i did not contribute to estimating $\hat{\tau}_j^{(-i)}(X_i)$).

The full sample is randomly split into two samples, S^A and S^B, each of which will be used both for training and predicting. Denote the trees resulting from these samples for each value of j by $\Pi_j^{S^A}$ and $\Pi_j^{S^B}$. Then consider the expectation of $\Gamma_{j,i}$ for a given partition

$$\zeta_{j,l}^A = E\left(\Gamma_{j,i}|X_i \in L_l\left(x;\Pi_j^{S^B}\right)\right),$$

$$\zeta_{j,l}^B = E\left(\Gamma_{j,i}|X_i \in L_l\left(x;\Pi_j^{S^A}\right)\right).$$

Let $L_l\left(x;\Pi_j\right)$ denote the lth element of the collection of leaves of the tree Π_j. The moments of all leaves are contained in $\zeta = \left(\zeta^A, \zeta^B\right)$. Recall that positive values of ζ point toward a local violation of LiM. Then a local violation of LiM can be tested with the following hypothesis test:

$$H_0 : \zeta_s \leq 0 \qquad \text{for all } s = 1, ..., p$$

$$H_1 : \zeta_s > 0 \qquad \text{for some } s = 1, ..., p,$$

where $p = |\zeta|$ is the number of sample splits. This means that $p = 2$, when splitting the sample into two samples, S^A and S^B.

Under the null hypothesis, an upper bound on the $(1-\alpha)$ quantile of $\sqrt{n}\left(\hat{\zeta}_j - \zeta_j\right)/\hat{\sigma}_j$ is enough for testing.:

$$T = \max_{1\leq s\leq p} \frac{\sqrt{n}\hat{\zeta}_j}{\hat{\sigma}_j} \leq \max_{1\leq s\leq p} \frac{\sqrt{n}\left(\hat{\zeta}_j - \zeta_j\right)}{\hat{\sigma}_j}.$$

Finally, the p-values should be Bonferroni corrected for multiple hypothesis testing.

Asymptotic results can be derived as in Farbmacher, Guber, and Klaassen (2022) using the results of Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins (2018).

C.3 Pseudo code of the LiM testing procedure

I build upon the procedure and code of Farbmacher, Guber, and Klaassen (2022) to establish a test for LiM. The pseudo code for the LiM test procedure is presented by Algorithm 1.

Algorithm 1 LiMtest

Input: n_s observations $(D_i, \widetilde{Z}_i, X_i)$ with $D_i \in \{0, 1, ..., J\}$ the treatment, $\widetilde{Z}_i$ the instrument indicator for the outer support of the instrument distribution, and X_i the covariates. The minimum leaf size is denoted k, and the significance level with α.

1: **for** $j = 0, 1, ..., J$ **do**
2: Construct the pseudo variable $Q_{j,i}$.
3: **for** both samples separately **do**
4: Obtain leave-one-out estimates $\hat{\mu}_j^{(-i)}(X_i)$ with a regression forest using outcome $Q_{j,i}$ and including covariates X_i.
5: Obtain leave-one-out estimates $\hat{\tau}_j^{(-i)}(X_i)$ with a causal forest using outcome $Q_{j,i}$ and including covariates X_i.
6: Construct the estimates $\hat{\Gamma}_{j,i}$ as in Equation (13) in Appendix C.2.
7: **end for**
8: Fit a CART tree on sample A using outcome $\hat{\Gamma}_{j,i}$, covariates X_i, minimal leaf size k, and apply cost complexity pruning.
9: **for** each leaf $l = 1, ..., l_{\max}$ **do**
10: Calculate the t-statistic $t_{j,l}^{(A)}$ over units $\hat{\Gamma}_{j,i}$ in sample A present in leaf l.
11: **if** $t_{j,l}^{(A)} > \Phi^{-1}(1 - 0.05/l_{\max})$ **then**
12: Calculate the t-statistic $t_{j,l}^{(B)}$ over units $\hat{\Gamma}_{j,i}$ in sample B present in leaf l and store the values in a vector T_{vec}.
13: **end if**
14: **end for**
15: Repeat lines 8-14 with the roles for samples A and B switched.
16: **if** $\max(T_{\text{vec}}) > \Phi^{-1}(1 - \alpha/|T_{\text{vec}}|)$ **then**
17: Reject the null hypothesis.
18: **end if**
19: **end for**

C.4 Simulation study - LiM test

In this section, we evaluate the performance of the LiM test through a simulation study. I conduct an empirical Monte Carlo study closely modeling the data from Card (1995). Specifically, I generate the instrumental variable $\widetilde{Z}$ using a binomial distribution with probabilities matching the mean observed. The control variables are generated are generated using a multivariate normal distribution, with means and covariance matrix derived from the empirical data. After generating the continuous control variables, I binarize them based on the mean. The sample size is 1500. For simplicity, I consider three different treatment levels, $D \in \{12, 13, 14\}$. Moreover, I introduce different complier types with specific probabilities of occurrence (type shares), as shown in Table 11. I consider two scenarios: one where LiM is valid, meaning no combined defiers are present, and another where LiM is locally violated due to the presence of defier types in the southern region, who change their treatment level from 13 to 12 when both instruments equal one. 1,000 simulation repetitions are performed.

Table 12 reports the rejection rates. The test successfully detects the violation for the combined defiers in the Southern region at the correct treatment level. The results reflect that the test is sensitive to the covariates included. As proxy variables might be chosen as most important variables for splitting, the tree structure with the maximum violation can only give an indication for the subgroup where LiM might be violated.

Table 11: This table presents an overview of the different response types and their respective shares. The simulation study considers two scenarios: one where LiM is valid, and another where LiM is locally violated due to the presence of combined defiers.

			LiM valid	LiM violated
	D^{00}	D^{11}	Type share	Type share
Combined compliers				
$cc_{12,13}$	12	13	15%	5%
$cc_{13,14}$	13	14	10%	10%
$cc_{12,14}$	12	14	5%	5%
Combined non-responders				
$cn_{12,12}$	12	12	25%	25%
$cn_{13,13}$	13	13	20%	20%
$cn_{14,14}$	14	14	25%	25%
Combined defiers				
$cd_{13,12}$ where $south = 1$	13	12	-	10%
			100%	100%

Table 12: This table presents the rejection results for the setting where LiM is valid and when LiM is violated.

	LiM valid		LiM violated	
	12 to 13	13 to 14	12 to 13	13 to 14
Rejection rate	0%	0%	100%	0%
First split at $south = 1$	-	-	94.8%	-

D Details on the model specifications

D.1 Specifications of the machine learners employed in Section 6.1.2

Lasso (Least Absolute Shrinkage and Selection Operator) introduced by Tibshirani (1996) is a regularization technique that performs both variable selection and regularization to enhance prediction accuracy. It includes a penalty term proportional to the absolute value of the coefficients, which helps in shrinking some coefficients to zero and thus performing feature selection. For Lasso, a fully saturated specification including all covariate interactions is considered, while raw controls are included in the other methods. The key hyperparameter is the regularization parameter, lambda. I use the *glmnet* library (version 4.1-8) in R to perform 5-fold cross-validation to select the optimal lambda.

Random Forest is an ensemble learning method used for classification and regression, introduced by Breiman (2001). It operates by constructing multiple decision trees during training and outputting the mode of the classes (classification) or mean prediction (regression) of the individual trees. Key hyperparameters include the number of trees, minimum node size, and the number of variables sampled at each split. The tuning process involved a grid search combined with 5-fold cross-validation. Specifically, the grid for the number of trees contained the values 500 and 1000. For the minimum node size, I consider values of 25, 50, and 100. The number of variables sampled at each split is varied with values of 2, 3, and 4. I use the *caret* package (version 6.0-94) and the *randomForest* package (version 4.7-1.1) to streamline the model training and hyperparameter tuning process.

Boosted Trees, specifically Gradient Boosting, is another ensemble learning technique that builds models sequentially, originally introduced by Friedman (2001). Each new model attempts to correct errors made by the previous models. The *gbm* library (version 2.1.9) in R is used for implementing Boosted Trees. I conduct a grid search on several key hyperparameters: the interaction depth, number of trees, and shrinkage rate. The grid for the interaction depth includes values of 1, 2, and 3. The number of trees is varied with values of 100, 200, and 300. For the shrinkage (learning rate), I consider values of 0.1, 0.05, and 0.01. 5-fold cross-validation is used to select the optimal combination of hyperparameter values.

In tuning the hyper parameters, for treatment and outcome models, RMSE (Root Mean Squared Error) was used to evaluate the model performance. Additionally, for each of the three methods (LASSO, Random Forest, and Boosted Trees), trimming of the predicted propensity scores was performed with a value of 0.01, such that they lie between 0.01 and 0.99.

Table 13: This table presents the mean RMSE over 25 splits for the machine learners used to obtain DML estimates in Card's (1995) study, as presented in Table 5, for selecting the best estimator among different machine learners.

	Mean RMSE for y	Mean RMSE for d	Mean RMSE for $\widetilde{Z}$
Lasso	0.42	1.96	0.37
Random forest	0.41	1.93	0.36
Boosted trees	0.41	1.92	0.36

D.2 Specifications of the machine learners employed in Section 6.2.2

The choices are similar to those outlined in Appendix D.1, with two exceptions: third-order covariate interactions are included for Lasso, and for random forests, only 100 and 500 trees are considered due to computation time.

Table 14: This table presents the mean RMSE over 5 splits for the machine learners used to obtain DML estimates in Angrist and Evans's (1998) study, as presented in Table 8, for selecting the best estimator among different machine learners.

	Mean RMSE for y	Mean RMSE for d	Mean RMSE for $\widetilde{Z}$
Panel A: Causal effect on labor income			
Lasso	4574.816	0.676	0.010
Random forest	4542.167	0.679	0.108
Boosted trees	4525.689	0.672	0.021
Panel B: Causal effect on hours worked per week			
Lasso	17.853	0.676	0.010
Random forest	17.855	0.679	0.108
Boosted trees	17.799	0.672	0.021
Panel C: Causal effect on weeks worked			
Lasso	21.30	0.68	0.01
Random forest	21.33	0.68	0.11
Boosted trees	21.25	0.67	0.02