Identifying Causal Effects of Nonbinary, Ordered Treatments using Multiple Instrumental Variables

Nadja van 't Hoff*

Abstract

This paper addresses the challenge of identifying causal effects of nonbinary, ordered treatments with multiple binary instruments. Next to presenting novel insights into the widely-applied two-stage least squares estimand, I show that a weighted average of local average treatment effects for combined complier populations is identified under the limited monotonicity assumption. This novel causal parameter has an intuitive interpretation, offering an appealing alternative to two-stage least squares. I employ recent advances in causal machine learning for estimation. I further demonstrate how causal forests can be used to detect local violations of the underlying limited monotonicity assumption. The methodology is applied to study the impact of community nurseries on child health outcomes.

Keywords: Nonbinary treatment, multiple instruments, average causal response, specification test.

JEL classification: C14, C21, C26.

This project was made possible through generous funding by the Independent Research Fund Denmark (90380031B). I would like to thank Giovanni Mellace for his valuable guidance, and my fellow PhD students for their constructive feedback. I further wish to thank Guido Imbens, Martin Huber, Phillip Heiler, and Leonard Goff for their insightful comments. I also acknowledge the participants of the "Machine Learning in Program Evaluation, High-Dimensionality, and Visualization Techniques" conference in 2023, the 2023 IAAE Conference, the seminar at Lund University in October 2023, and the DGPE workshop 2023.

*Department of Economics, University of Southern Denmark.

1 Introduction

A large amount of literature is devoted to estimating causal effects of a binary treatment using a single instrument. The problem of identifying the causal effect of a nonbinary, ordered treatment using multiple instruments remains relatively underexplored. Nevertheless, one can think of numerous examples where the treatment might attain integer values and go beyond the binary case. A typical example is the effect of schooling, which is often estimated with a binary variable for school attendance or completion, even though it might be more interesting to consider the years of schooling instead. Moreover, multiple instruments might be available, e.g., the quarter of birth (Angrist and Imbens, 1995), the distance to the school, and the local labor market conditions (see for instance Carneiro et al., 2011). Yet generalizing binary treatment results to this context is complex. Simply binarizing a treatment is discouraged, as it can violate the exclusion restriction (Andresen and Huber, 2021) and lead to inconsistent estimates (Angrist and Imbens, 1995), where only the sign but not the magnitude of the estimates can be trusted. Not to mention that binarizing the treatment alters the interpretation of the parameter and can be sensitive to the choice of the cut-off point.

Seminal work by Angrist and Imbens (1995) advanced our understanding of handling discrete, ordered treatments with multiple discrete instruments using two-stage least squares (TSLS). They revealed that TSLS estimates correspond to a weighted average of average causal responses (ACRs), where each ACR in itself represents a weighted average of causal effects of one-level treatment changes for specific response subpopulations. Simply put, TSLS estimates represent a weighted average of weighted averages of local causal responses. While the TSLS approach has become common practice in this setting, it comes with several limitations. Firstly, the TSLS estimand is hard to interpret. Secondly, the TSLS estimand often relies on a strict form of the monotonicity assumption to have a meaningful causal interpretation.

When treatment effects are heterogeneous, imposing a monotonicity assumption is essential to obtain a causally-interpretable effect. This assumption restricts the direction in which the potential treatment status changes for a given change in the instrument values. This translates to ruling out the presence of certain response types in the population and hence imposes restrictions on the possible choice mechanisms. Angrist and Imbens (1995) derived their results under the rather restrictive Imbens and Angrist monotonicity (IAM) assumption. In the case of multiple instruments, this assumption imposes a direction for every possible change in instrument values, therefore heavily constraining choice behavior. Motivated by this observation, Mogstad et al. (2021) introduce the weaker partial monotonicity (PM) assumption in the setting of a binary treatment and multiple instruments. Within the same setting, van 't Hoff et al. (2023) introduce the limited monotonicity (LiM) assumption, which allows for the presence of many different response types and hence imposes even fewer restrictions on the choice mechanisms. The fewer restrictions are imposed, the larger the possible response population, which increases the credibility of this assumption and makes it more broadly applicable. The present study extends the notions of both PM and LiM to the framework of a nonbinary, ordered treatment. Next to the advantage of rich choice heterogeneity, I demonstrate that LiM allows for simplifying the multiple instrument problem into a single instrument problem by aggregating complier types into groups.

Equipped with the extended versions of monotonicity, a key contribution of this study is a general TSLS identification result that flexibly accommodates the different monotonicity assumptions. The challenge with TSLS is navigating a landscape of suboptimal options. On the one hand, IAM ensures positive weights, but imposes many constraints on choice behavior. On the other hand, PM imposes fewer constraints on choice behavior, but it potentially introduces negative weights. In both cases, the TSLS estimator converges to a weighted average of weighted averages of ACRs over the response types that remain under the imposed monotonicity assumption, with weights that complicate a straightforward interpretation of the TSLS estimates.

The main theoretical contribution of this study is a novel causal parameter, the *combined compliers average causal response* (CC-ACR), which is identified under the relatively mild LiM assumption and has an intuitive interpretation. Following van 't Hoff et al. (2023), combined compliers are defined as those units whose treatment level changes in the same direction when all instrument values shift simultaneously from zero to one. The weights of the proposed CC-ACR parameter are positive by construction and proportional to the shares of the combined complier types along the specific treatment margin, which arguably are ideal weights. The CC-ACR provides valuable insights as it addresses the simple and informative question: "What is the average causal response for a one-level increase in the treatment variable for combined compliers?" This parameter encompasses a substantial complier population and is robust against various defier types. Altogether, the CC-ACR parameter successfully tackles the two challenges associated with the widely used TSLS methodology, as it provides a clearer interpretation under the comparably mild LiM assumption.

Another pivotal contribution of this paper is a stochastic dominance test for detecting violations of the LiM assumption. LiM implies positive weights for the CC-ACR parameter, which I show to be equivalent to the condition that the cumulative distribution functions (CDFs) of the treatment conditional on specific instrument values do not intersect. This necessary (though not sufficient) condition is verifiable with the data. If violations of LiM occur in specific subgroups, a global test might fail to detect a violation of this assumption as violations of LiM might average out when considering the full sample. Therefore, expanding upon Farbmacher et al. (2022), I demonstrate how causal forests can be used to detect such local violations. In essence, the testing procedure boils down to verifying the sign of conditional average treatment effects (CATEs) within the leaves of a regression tree, with a positive sign signaling a violation of LiM. This approach offers the advantage of data-driven subgroup formation in the covariate space and, in addition, has larger power than alternative tests when the degree of violation varies across subgroups.

To illustrate the proposed methods, I revisit the application by Attanasio et al. (2013), who study the *Hogares Comunitarios* (HC) program. The HC program, a crucial policy implemented by the Colombian government, establishes community nurseries and aims at improving nutrition during early childhood. This paper seeks to identify and estimate the average causal effect of an additional month enrolled in the HC program on a child's height, a key indicator for nutritional status. The treatment corresponds to the number of months that a child attends the program, meaning that it is discrete and ordered. To address potential confounding factors influencing both HC attendance and nutritional status, I employ a diverse set of instruments that measure time and money costs associated with HC attendance. These instruments include the distance to the HC nursery, the HC nursery's capacity, and the fee paid to the "madre comunitaria". Then, the combined compliers represent children whose program participation would be extended if the program were simultaneously made more affordable, easily accessible, and more proximate to their homes, and the CC-ACR parameter addresses the question: "What is the average monthly causal effect of the HC program on a child's height for these combined compliers?"

Following Attanasio et al. (2013), I assume that the instruments are exogenous, given certain covariates such as household and neighborhood characteristics. I establish an identification result for the CC-ACR under this condition. To ensure a clear interpretation of CC-ACR estimates when instrument validity holds only conditionally, the characteristics are flexibly controlled for. For estimation, I view the CC-ACR as a ratio of average treatment effects and leverage recent developments in causal machine learning. Specifically, I use generalized random forests (GRF) (Athey et al., 2019) and double/debiased machine learning (DML) (Chernozhukov et al., 2018). While Attanasio et al. (2013) find that participation in the HC program has a significant positive effect on a child's height, I find no significant effect of HC attendance on a child's height. This finding suggests the program's need for re-evaluation and potential improvements to better achieve its goal of enhancing the health outcomes of underprivileged children.

In the HC application, LiM seems more reasonable an assumption than PM. For instance, unlike PM, LiM does not rule out parents who perceive a high capacity of a nursery as a negative signal, as long as these parents can be incentivized to extend their child's program participation by reducing the distance to the nursery or by reducing the associated fees. A visual examination of global stochastic dominance does not suggest a violation of LiM. However, a potential problem is that LiM might be violated only for certain subgroups in the data. For example, parents who perceive a high capacity of a nursery as a negative signal might be wealthier or have other childcare options available. Using the proposed LiM test, I successfully detect a violation of LiM within a subgroup of the data. The violation is detected for households that are likely located at the higher end of the wealth distribution. This means that the CC-ACR estimates should be interpreted with caution, and the TSLS estimates even more so.

The remainder of this paper is organized as follows: Section 2 provides an overview of the literature. Section 3 provides the overall framework and assumptions, as well as the CC-ACR and TSLS identification results. Section 4 provides a procedure for detecting violations of the LiM assumption, and Section 5 establishes guidelines for estimation. This is followed by Section 6, which reports the empirical findings of the effect of the HC program using the methods outlined in the previous sections. Section 7 provides a discussion and avenues for future research. For brevity, some results, including an extension to unordered treatments, are presented in the appendix.

2 Literature review

This paper contributes to the instrumental variables literature in two key areas. First, it contributes to the identification of causal effects. The foundation of the local average treatment effect (LATE) framework was established by Imbens and Angrist (1994) and Angrist et al. (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 et al. (2021) provide an identification result for TSLS under PM, and van 't Hoff et al. (2023) show that the LATE for the combined compliers is identified under LiM. Frölich (2007) extends the LATE framework to include covariates nonparametrically. My findings complement those of Frölich (2007) who separately considers a nonbinary, ordered treatment with a single instrument, or a binary treatment with multiple instruments, while my results consider the setting with nonbinary, 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 et al. (2006) considers 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 et al.'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.

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 et al. (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 and Pearl, 1997; Kitagawa, 2015; Mourifié and Wan, 2017; Huber and Mellace, 2015; Frandsen et al., 2023; Carr and Kitagawa, 2021). Farbmacher et al. (2022) builds 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 (2020) 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 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: Independence and exclusion

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

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 the HC application, the independence assumption posits that factors like distance to an HC nursery do not influence a child's nutritional status except through the HC program. 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. 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 an HC nursery influences HC attendance for some children.

The limited monotonicity (LiM) assumption was initially introduced by van 't Hoff et al. (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 the HC application, this implies that the number of months spent in the HC program while residing close to it, experiencing low fees, and having access to a nursery with high capacity, is at least as large as the number of months when residing far, facing high fees, and having access to a nursery with low capacity.

LiM is generally weaker than other monotonicity assumptions introduced in literature, such as the Imbens and Angrist monotonicity (IAM) assumption (Imbens and Angrist, 1994) and the partial monotonicity (PM) assumption (Mogstad et al., 2021). 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 $J^{2^K} = 3^{2^1} = 9$ 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).

The number of initial response types increases rapidly with the number of treatment levels. For nonbinary, ordered 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. Consider the context of the HC program as an example. Some parents might enroll their child for zero months in the program when they live far away from an HC nursery, while their neighbors might enroll their child for one month. The baseline treatment level equals zero for the first child and it equals one for the neighboring child. Suppose now that these two families are moved closer to the HC nursery. While the first parents might now enroll their child for two months instead of zero months, their neighbors might enroll their child for two months instead of one month. This means that the compliance intensity is two levels for the first child while it is one level for the neighboring child.

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 6, for a

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			

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.[1] Under IAM, there are only 14 response types that remain.[2] 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.

Next 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 $\mathscr{Z} = \{(1,1),(1,0),(0,1),(0,0)\}$, the two intermediate treatment states, D^{10} and D^{01}, are not restricted. Therefore, aggregating the initial response types into response type groups is straightforward. With two instruments, I define as combined compliers, denoted as $cc_{D^{11},D^{00}}$, 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 $cd_{D^{11},D^{00}}$, those types who have $D^{11} < D^{00}$, and as combined non-responders, denoted as $cn_{D^{11},D^{00}}$, those types who have

[1] Note that there is only one ordering of PM that is consistent with the data. Define $\bar{D}^{z_1,z_2} = \frac{1}{\sum_i z_{1,i} \cdot z_{2,i}} \sum_i z_{1,i} \cdot z_{2,i} \cdot D_i$. Then, in the HC application, $\bar{D}^{000} = 5.4$, $\bar{D}^{001} = 7.9$, $\bar{D}^{100} = 9.4$, $\bar{D}^{010} = 10.4$, $\bar{D}^{011} = 12.2$, $\bar{D}^{110} = 15.0$, $\bar{D}^{101} = 15.2$, and $\bar{D}^{111} = 16.3$, implying the following ordering: $E(D|Z = 0,0,0) < E(D|Z = 0,0,1) < E(D|Z = 1,0,0) < E(D|Z = 0,1,0) < E(D|Z = 0,1,1) < E(D|Z = 1,1,0) < E(D|Z = 1,0,1) < E(D|Z = 1,1,1)$, which is nested with LiM.

[2] For a detailed comparison of the three monotonicity assumptions when the treatment is binary see van 't Hoff et al. (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 $cc_{D^{11},D^{00}}$, combined non-responders by $cn_{D^{11},D^{00}}$, and combined defiers by $cd_{D^{11},D^{00}}$.

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	

$D^{11} = D^{00}$. It is important to highlight that aggregating the groups in this way is not possible under PM or IAM.

Illustratively, consider the nine initial types in Table 2, extracted from Table 6 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 6 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, $D^{1...1...1}$ and $D^{0...0...0}$, independently of the number of instruments. Notably, combined defier types with $cd_{a,b}$ where $a < b$ are ruled out by the LiM assumption, but all other defier types are not.

3.2 The combined compliers ACR

Theorem 1 provides my main identification result, namely the CC-ACR, which has a straightforward interpretation and is derived under the generally weaker LiM assumption.

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

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

Proof in Appendix B.1.

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 HC application, it provides a weighted average of causal responses for those children who extend their enrollment duration in the program when all instrument values shift from zero to one. Recall that all instrument values equaling one means that children reside close to the nursery, experience low fees, and have access to a nursery with high capacity. To clarify the meaning of Theorem 1, note that $E(Y^3 - Y^1|cc_{1,3})$ measures the average effect on a child's height when attending the HC nursery for three months instead of one month, for those children that change their treatment status correspondingly in response to this change in instrument values. This average effect is weighted by the probability of belonging to this particular complier type, represented by $P(T = cc_{1,3})$, offering the advantage of weights that are proportional to the response type group size. Further note that $E(Y^3 - Y^1|cc_{1,3})$ gives the effect of $(3 - 1) = 2$ additional months in the HC program. This is reflected in scaling the difference in outcomes, $Y^3 - Y^1$, by the treatment level difference, $(l - k) = (3 - 1) = 2$ months.

Theorem 1 simplifies if treatment effects are linear in the sense that the impact of increasing HC enrollment by one month is equivalent whether it is, for example, from one to two months or from eight to nine months. In this case, the CC-ACR can be interpreted as the average effect for a one-level increase among the combined complier population: $E(Y^j - Y^{j-1}|T \in cc)$, 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

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

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

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

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 HC application as an example, one might worry about factors that influence a family's surroundings as well as a child's nutritional status. The distance instrument, for instance, might be correlated with various household and neighborhood characteristics. In this case, Assumption 1 has to be adjusted such that the instruments are approximately randomly assigned conditional on the household and neighborhood characteristics.

Assumption 1C: Unconfoundedness and exclusion

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

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

$\text{Supp}(X|Z_1 = 1, Z_2 = 1, ..., Z_K = 1) = \text{Supp}(X|Z_1 = 0, Z_2 = 0, ..., Z_K = 0).$

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

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

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 $\widetilde{Z} = Z_1 = Z_2 = ... = Z_K$. 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 et al. (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 nonbinary, ordered 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 $\mathscr{Z} = \{z_0, ..., z_l, ..., z_m\}$, 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 $\mathscr{Z} = \{z_0, ..., z_l, ..., z_m\}$, 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 of $\iota_{m,m-1}$ being always smaller 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.

3.5 Alternative representation of the CC-ACR

Theorem 1 introduces 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 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]

[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 $J = 3$, meaning that there are six possible 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.

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.

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.

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

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. For instance, Attanasio et al. (2013) explore the impact of a child's exposure to the HC program, quantified as the ratio of months enrolled in the program to the child's age in months, on their height. Theorem 2 provides the results for the continuous treatment setting.

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 $D^{0...0...0}$.

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. 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 $\widetilde{Z} = Z_1 = Z_2 = ... = Z_K$. 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 $i = 1, ..., n$, 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

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

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

[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.

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 et al. (2022) (see Appendix C.2 for a detailed description). First, a causal forest (Wager and 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 et al., 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] When a violation of LiM is detected, a simple solution consists of only considering a subset of the instruments for which no violation of LiM is detected. However, it is important to acknowledge that this approach alters the interpretation of the parameter and results in estimating effects for a smaller combined complier population.

[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 et al. (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.

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 increases the share of combined compliers considered, possibly resulting in a variance reduction. For a more detailed discussion, refer to van 't Hoff et al. (2023).

5.2 Estimation with covariates

In Section 3.3, I established the identification result of the CC-ACR under the assumption that conditional independence holds, assuming adequate overlap. While it may be tempting to incorporate these covariates linearly into the TSLS approach, linear inclusion of the covariates can introduce interpretation complexities already in the context of binary treatments (Blandhol et al., 2022; Słoczyński, 2020). Blandhol et al. (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.

Instead of linearly including the covariates and using the TSLS estimator, this paper proposes an approach for estimating Equation (4) which flexibly incorporates covariates. An avenue to estimation views Equation (4) as a ratio of two average treatment effects (ATEs). Estimation then boils down to estimating the numerator and denominator of Equation (4) separately. In this context, causal machine learning methods are exceptionally valuable as they are not prone to the curse of dimensionality.[11] These nonparametric methods estimate $\sqrt{n}$-consistent average

[11] In case of few discrete covariates, kernel regression like Nadaraya-Watson regression offers an interesting

treatment effects while flexibly incorporating high-dimensional covariates. For instance, the generalized random forest (GRF) (Athey et al., 2019) can be employed to estimate the numerator and denominator separately. This method imposes a partially linear model and relies on GMM estimation with forest-constructed weights. Alternatively, a partially linear model within the double/debiased machine learning (DML) framework (Chernozhukov et al., 2018) can be used to estimate the two components of Equation (4). The nuisance components of DML can be estimated, for example, with random forests (Breiman, 2001).[12] [13] Standard errors can be obtained by bootstrapping, or alternatively the delta method.

In summary, the proposed approach views the estimator as a ratio of two ATEs and uses causal machine learning techniques for estimation.

6 Empirical application to the HC program

In this section, I implement the result to explore the causal effect of the HC program on a child's height. The HC program is an important policy aimed at improving nutrition at a young age, launched by the Colombian government in 1984. Its aim is crucial as malnutrition not only adversely affects health (Martins et al., 2011), but also has long-term consequences on educational attainment and productivity (Prado and Dewey, 2014). The HC initiative establishes local nurseries, led by a "madre comunitaria", to provide care and meals for children in need, aiming to meet 70% of their daily caloric needs (González Ramírez and Durán, 2012). Each nursery takes care of up to 15 children aged zero to six. With nearly 80,000 HC nurseries serving around a million children in 2011, the program's reach is extensive, and given its substantial allocation of 0.3% of Colombia's GDP (González Ramírez and Durán, 2012), it is pivotal to assess the program's effectiveness at improving child health outcomes.

alternative to estimation. In case of high-dimensional covariates, dimensionality reduction techniques can be combined with kernel regression, e.g. principal component analysis (Hotelling, 1933) or selecting a subset of most relevant variables using instrumental forests (Athey et al., 2019). However, caution with these techniques is required as the former might alter the interpretation of the CC-ACR, whereas the latter potentially introduces omitted variable bias.

[12] Chernozhukov et al. (2018) finds that the specific machine learning method (XGBoost, random forests, Lasso, amongst others) used for estimating the nuisance function does not qualitatively change the conclusion of the considered applications. The machine learning methods all provide good approximations to the true functions.

[13] A potential concern is that the semiparametric efficiency bound is probably not attained, though investigating the efficiency of this approach extends beyond the scope of this paper.

6.1 Data

I use data from the Familias en Acción survey, which was previously used by Attanasio et al. (2013), to study the effect of the HC program on a child's height, instrumenting with measures that capture the costs of program attendance. In this section, I highlight the most relevant aspects regarding the data set, referring the interested reader to Attanasio et al. (2013) for a more detailed discussion.

The prevailing treatment considered in the present paper is the number of months that a child attends the HC program, $D \in \{0, 1, ..., 48\}$. I also consider the continuous treatment variable that captures a child's exposure to the HC program measured as the number of months in the program divided by the child's age in months. The small number of children who participated in the HC program for more than 48 months have been excluded from the sample. The outcome variable considered is height, standardized by age and gender, which provides insights into the nutritional status of the children. This study makes use of three instrumental variables: (1) the household's travel time to the nearest HC nursery[14], (2) the median fee of the HC nursery in the municipality as of December 2003, and (3) the capacity of the local HC nursery, which is measured as the ratio of the available spots in a town compared to the total number of eligible children aged between two and six years.[15] In order to facilitate the analysis, I binarize all three instruments based on their respective median values. Therefore, the combined complier population are those children whose enrollment in the program would be extended if the instruments were shifted from below the median value to above the median value.

Given the concern with instrument validity raised in Attanasio et al. (2013), I follow their approach and include a comprehensive set of controls in the analysis. These controls encompass factors like distances to key facilities like schools, health centers, and the town hall. In addition, town-level variables like the proportion of children with health insurance and the number of hospitals are included, as well as household-level characteristics such as the educational level of the mother and head of household, the height of the mother, the gender of the child, and the birth order of the child. In contrast to Attanasio et al. (2013), I exclude certain continuous town-level variables, including the percentage of households with sewage connection, the percentage with pipe water, and the altitude, due to poor overlap. To account for potential systematic variations across different survey waves and geographical regions, fixed effects are introduced. These fixed effects specifically pertain to the survey waves (wave 1, wave 2, and wave 3) as well as the

[14] Specifically, it refers to the distance to the attended HC nursery if a child attends and to the closest HC nursery if a child does not attend.

[15] This means that there are $(J+1)^{2^K} = 49^8$ different initial response types. LiM always rules out 25% of the response types, meaning that $0.75 \cdot 49^8$ response types are allowed for. The combined complier population consists of $0.5 \cdot 49^8$ types.

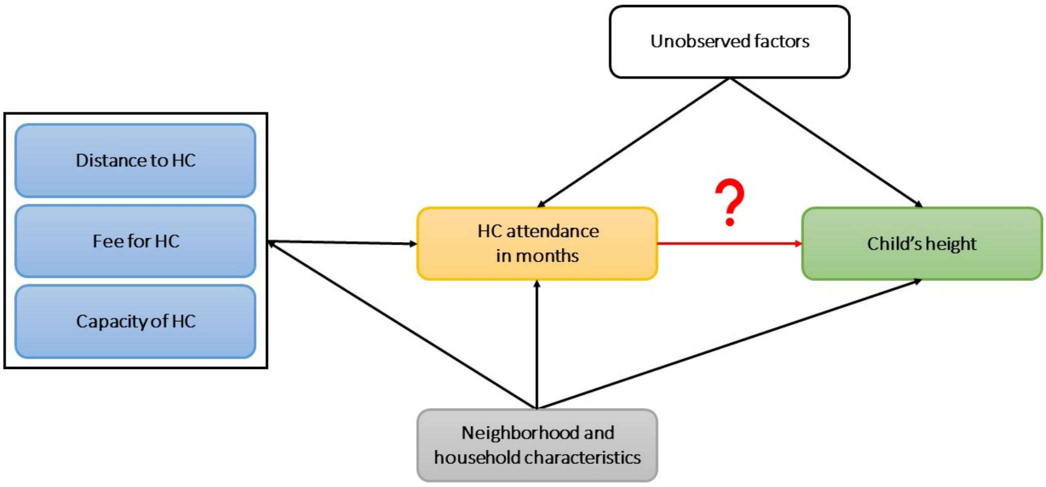

Figure 1: DAG for the HC application. The goal is to estimate the average causal effect of an additional month in the HC application, indicated by the red arrow and question mark. Unobserved factors (e.g., culture) that influence both the treatment and the outcome lead to endogeneity in the treatment variable. Three instrumental variables enable identification of the treatment effect. These are assumed to be valid after controlling for neighborhood and household characteristics.

various geographic regions. Table 7, presented in Appendix D, contrasts the mean values of these characteristics under two distinct scenarios: when all instrument values are zero and when all instrument values are one. Indeed, for the majority of covariates, the averages at the outer support of the instrument distribution differ significantly, indicating that their inclusion might be necessary for the validity of the instruments. Next to instrument validity, including covariates in estimation might offer additional advantages. By capturing some of the variability in the outcome, covariates might increase precision. Moreover, covariates can account for imbalances between the treated and non-treated populations, potentially increasing the credibility of the results for other populations. A visual overview of the HC application is given by the Directed Acyclic Graph (DAG) in Figure 1.

6.2 Analysis of the causal effect of HC nurseries on a child's height

In this section, I present the empirical findings from implementing the TSLS methodology as well as the proposed CC-ACR methodology. It is important to note that differences in estimates can be due to differences in underlying complier populations, underlying assumptions, potential violations of the partial monotonicity assumption, or the different weighting schemes.

The results of using the standard ordinary least squares (OLS) and TSLS approaches are presented in Table 4. The TSLS specification is saturated in the instruments in the first stage. OLS is expected to be downward biased because poorer families might be more likely to enroll their children in the HC program and these children attain generally a lower height due to malnutrition. The OLS estimates in Columns (1) and (2) provide some indication that there is selection into treatment, highlighting the need for instruments to obtain an effect of the HC program on height with a causal interpretation. The standard TSLS estimates in Columns (3) and (4) show a non-significant effect of HC attendance on height. Columns (5) and (6) in Table 4 contain the CC-ACR estimates without controls and with linear inclusion of controls, respectively. The interpretations of the TSLS estimates in Column (4) and the CC-ACR estimates in Column (5) are problematic as covariates are included linearly, potentially including the effects of never-takers and always-takers, which might contaminate the estimated averages of combined complier effects (see Section 5.2).

I now shift my focus to methods that offer greater flexibility in accounting for confounding factors. The results, employing causal machine learning techniques, are presented in Columns (7) and (8) of Table 4. Standard errors are derived through cluster bootstrapping at the municipality level. The CC-ACR estimates in Column (7) are obtained using the GRF method (Athey et al., 2019) for the numerator and denominator in Equation (4) separately. Similarly, the CC-ACR estimate in Column (8) is obtained using the DML framework (Chernozhukov et al., 2018) with random forests (Breiman, 2001) for estimating the nuisance components.[16] The model specifications for Columns (7) and (8) are provided in Appendix E.1.1 and Appendix E.1.2, respectively.

Across the estimation methods, the CC-ACR estimates indicate no significant effect of the HC program on a child's height. Yet, for illustration purposes, I would like to comment on the interpretation. Assuming linearity, each additional month of attending a HC nursery corresponds to a specific average change in height for the population of combined compliers. Conversely, for non-linear effects, the interpretation of the estimates is the average monthly height change considering a 48-month enrollment period for the combined compliers. For instance, an estimate of 0.001 (0.013) implies that a one-month increase in HC attendance is associated with an average height increase of approximately 0.1% (1.3%) of one standard deviation in height. Over four years of HC attendance, this translates to a potential average increase of 4.8% (62.4%) of one standard deviation in height. This is a local effect for the combined compliers, so those types who would increase their treatment uptake when all instruments are being changed from zero to one.

Next, I consider the continuous treatment variable that captures a child's exposure to the

[16] Alternative methods for estimating the nuisance components, like XGBoost Chen and Guestrin (2016) and Lasso Tibshirani (1996), produce similar estimates, though with larger standard errors.

Table 4: This table presents the OLS estimates, conventional TSLS estimates, and CC-ACR estimates of the average causal effect of an additional month in the HC program on a child's standardized height.

	$\hat{\beta}_{\text{OLS}}$		$\hat{\beta}_{\text{TSLS}}$		$\hat{\beta}_{\text{CC-ACR}}$		$\hat{\beta}_{\text{CC-ACR-GRF}}$	$\hat{\beta}_{\text{CC-ACR-DML}}$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
HC attendance	-0.003	-0.003^{*}	-0.001	0.016	0.004	0.035	0.001	0.013
(Std. err.)	(0.002)	(0.002)	(0.009)	(0.015)	(0.009)	(0.025)	(0.009)	(0.017)
Observations	5,574	5,574	5,574	5,574	2,681	2,681	2,681	2,681
% observations	100%	100%	100%	100%	48%	48%	48%	48%
Covariates	no	linear	no	linear	no	linear	flexible	flexible

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

Standard errors are clustered at the municipality level.

Cluster bootstrapped standard errors (100 repetitions) for Columns (7) $\hat{\beta}_{\text{CC-ACR-GRF}}$ and (8) $\hat{\beta}_{\text{CC-ACR-DML}}$.

Table 5: This table considers the continuous treatment variable which measures the exposure to the HC program, and table presents the OLS estimates, conventional TSLS estimates, and CC-ACR estimates of the causal effect on a child's standardized height.

	$\hat{\beta}_{\text{OLS}}$		$\hat{\beta}_{\text{TSLS}}$		$\hat{\beta}_{\text{CC-ACR}}$		$\hat{\beta}_{\text{CC-ACR-GRF}}$	$\hat{\beta}_{\text{CC-ACR-DML}}$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
HC exposure	-0.119	-0.151	-0.058	0.089	0.212	1.685	0.026	0.660
(Std. err.)	(0.125)	(0.101)	(0.506)	(0.077)	(0.457)	(1.165)	(0.480)	(0.872)
Observations	5,574	5,574	5,574	5,574	2,681	2,681	2,681	2,681
% observations	100%	100%	100%	100%	48%	48%	48%	48%
Covariates	no	linear	no	linear	no	linear	flexible	flexible

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

Standard errors are clustered at the municipality level.

Cluster bootstrapped standard errors (100 repetitions) for Columns (7) $\hat{\beta}_{\text{CC-ACR-GRF}}$ and (8) $\hat{\beta}_{\text{CC-ACR-DML}}$.

HC program measured as the number of months in the program divided by the child's age in months (refer to Section 3.7 for the identification result). Again, I leverage the causal machine learning methods for estimation. Detailed specifications for the parameters of the GRF method are presented in Appendix E.1.4, while the DML specifications are the same as those outlined in Appendix E.1.2. The estimates in Table 5 vary in magnitude, but none of them are statistically significant at the 5% level. An estimate of 0.026 suggests that a child that is exposed to the program for 100% of the 48 months is on average 2.6% of one standard deviation taller in height.

In conclusion, the findings closely align: There is no statistically significant causal effect of the HC program on a child's height. My results differ from Attanasio et al.'s (2013) results, which implied that a child aged between two and six years that attends an HC their entire life will be 88.5% of one standard deviation of height taller, and a meta-analysis by Bhutta et al. (2008), who stated an increase of 41% of the standard deviation. However, differences in estimation strategy, underlying assumptions and interpretation of the estimates should be considered. The focus of the next sections will lie on the discrete, ordered treatment as measured by the number of months enrolled in the HC program.

6.3 Heterogeneity analysis

To study the heterogeneity of treatment effects, instrumental forests (Athey et al., 2019) are well-suited, as they enable the estimation of $\beta^{(-i)}_{\text{CC-ACR}}(X_i)$ of Equation (3), where the superscript $(-i)$ denotes out-of-bag estimates. Figure 2 shows that there is some variation in these conditional CC-ACR estimates, with some of the estimates attaining negative values, which indicates a negative effect of the HC program on height for some children.

After having established that there appears to be some variation in treatment effects, it is of interest to study which variables cause this heterogeneity. Partial dependence plots (Friedman, 2001) can be informative on univariate heterogeneity based on observable characteristics. Thus, they can help explain the mechanisms behind the observed heterogeneity in the estimated $\beta_{\text{CC-ACR}}(X_i)$. To study how the causal effect of the HC program on a child's height varies with, for example, the mother's age, the $\beta_{\text{CC-ACR}}(X_i)$ is predicted using the instrumental forest while varying the age and fixing the values of the remaining covariates to their median value. Figure 3 depicts the resulting plot. It shows that the HC program has a positive effect on a child's height if its mother is aged below 40. For children whose mother is aged above 40 years, the program seems to have a small negative effect. A possible explanation could be that older parents are more financially stable and better able to provide nutritious meals for their children compared to the meals provided by the HC nursery. However, this observation is not significant and speculative, therefore warranting further investigation.

Next, I study heterogeneity along two dimensions. Figure 4 provides an overview of the estimated $\beta_{\text{CC-ACR}}(X_i)$ while varying both the mother's age and her educational attainment (secondary school completion or more). Surprisingly, the HC program appears to have a potentially adverse effect on children when their mothers are older than 40 and lack secondary school completion, whereas it appears to be most beneficial for children whose mothers are aged between 26 and 30 and completed secondary education or higher. However, none of these estimates are statistically significant at the 5% level.

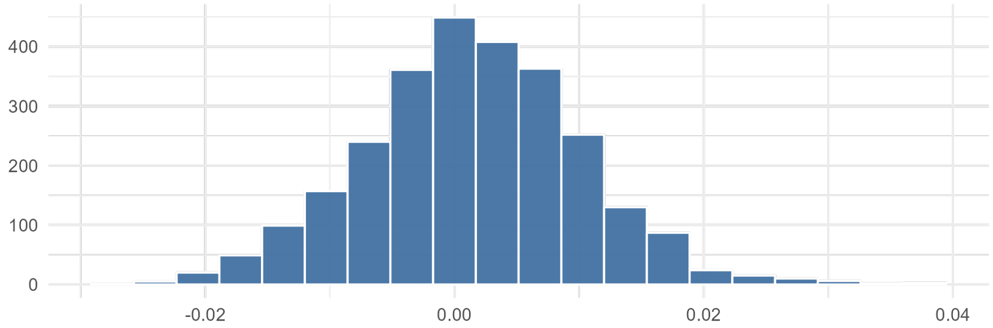

Figure 2: Frequency of out-of-bag conditional $\beta_{\text{CC-ACR}}$ estimates, as estimated with the instrumental forest, which shows that there is some variation in these estimates.

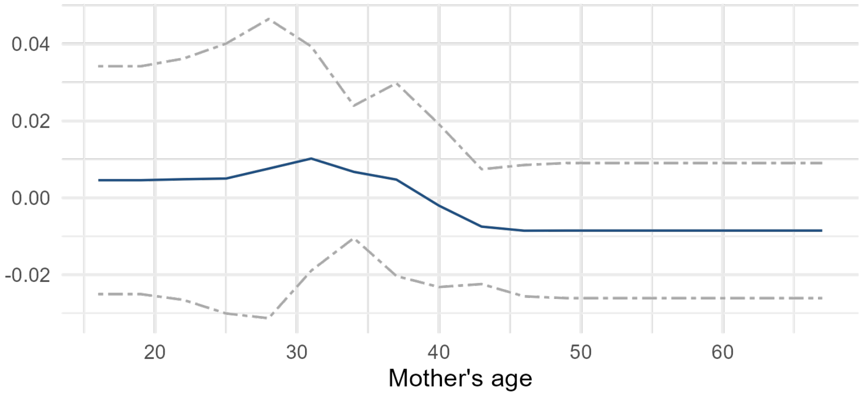

Figure 3: Partial dependence plot for univariate heterogeneity along the mother's age, as estimated with the instrumental forest. 95% confidence intervals are indicated by the dashed lines.

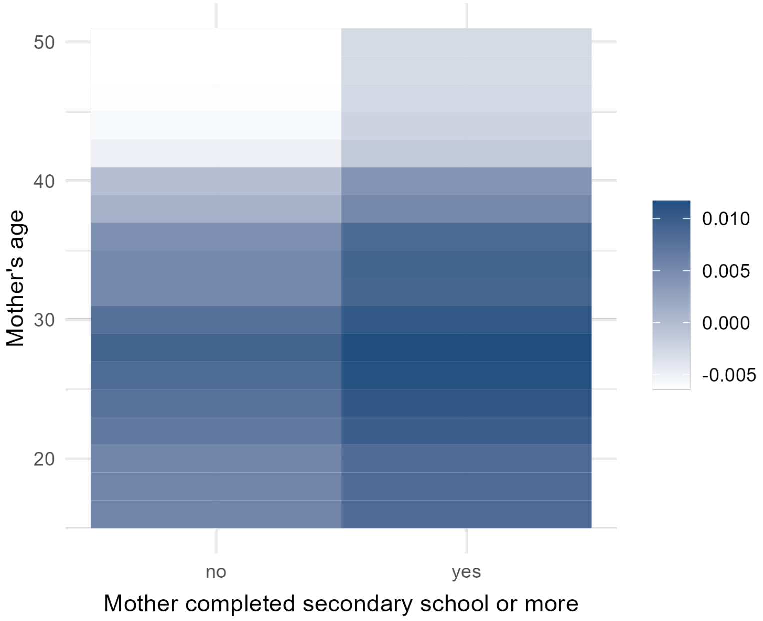

Figure 4: Heatmap depicting bivariate heterogeneity along mother's age and educational attainment (secondary school completion or more), as estimated with the instrumental forest. The estimated conditional CC-ACRs are depicted in the plot.

6.4 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 outer support of the instrument values, $P(D < j|Z_1 = Z_2 = ... = Z_K = 0)$ and $P(D < j|Z_1 = Z_2 = ... = Z_K = 1)$. Figure 5 depicts the two CDFs and Figure 6 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 month j, the difference is the share of the children whose enrollment in the HC program extends from j months to j months or more in response to the shift in instrument values. Unsurprisingly, more weight is placed on treatment margins at the lower end of the causal response function. Convincingly, this indicates that the instruments disproportionately affect the enrollment for children who would only have been enrolled for a brief period in the absence of the instruments.

6.5 LiM test

LiM can be violated if there are defiers with respect to one of the instruments that cannot be pushed towards compliance by some other instrument. This means that if there are, for example, defiers with respect to the capacity instrument that cannot be pushed towards compliance by offering a lower fee or a reduced distance to the nursery, then LiM is violated. A visual inspection of Figure 5 does not indicate a global violation of LiM, as the CDFs do not cross.

As violations of LiM might occur within subgroups of the observed characteristics, I implement the local LiM test as formulated in Section 4.2. To implement the proposed testing procedure, I adapt the R package developed by Farbmacher et al. (2022) to suit my specific needs.[17] I include a variety of control variables that likely influence parents' decision to enroll their child in the program as well as the outcome. Specifically, the educational attainment of both the mother and head of household, the age of both the mother and head of household, the child's gender, the child's birth order within the family, the food price index, and the female urban and rural wages. These controls are valid as they are likely unaffected by the instruments and treatment. It is important to acknowledge 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.

[17] 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 100.

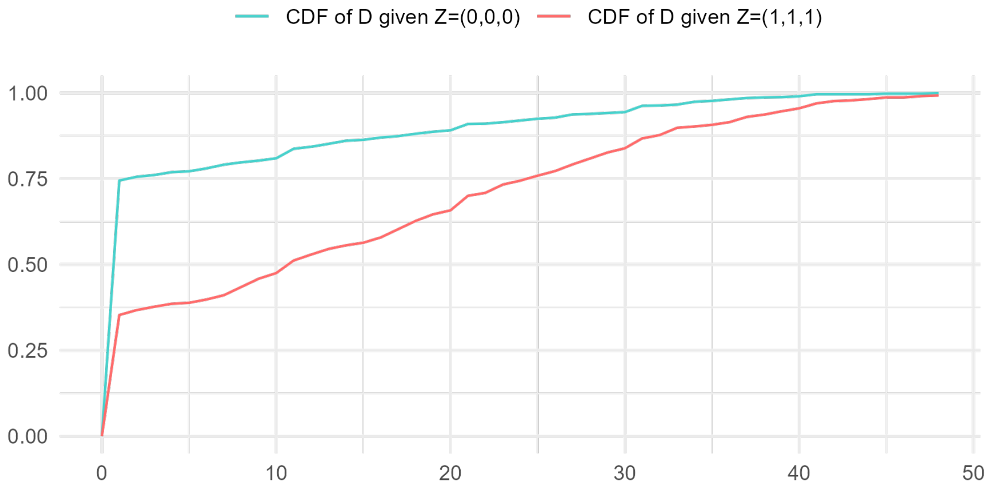

Figure 5: CDFs of the treatment conditional on the outer support of the instrument values when using all three instruments. The CDFs do not cross, indicating that the necessary condition for LiM holds at all treatment levels and no violation of the LiM assumption is detected visually in the full sample.

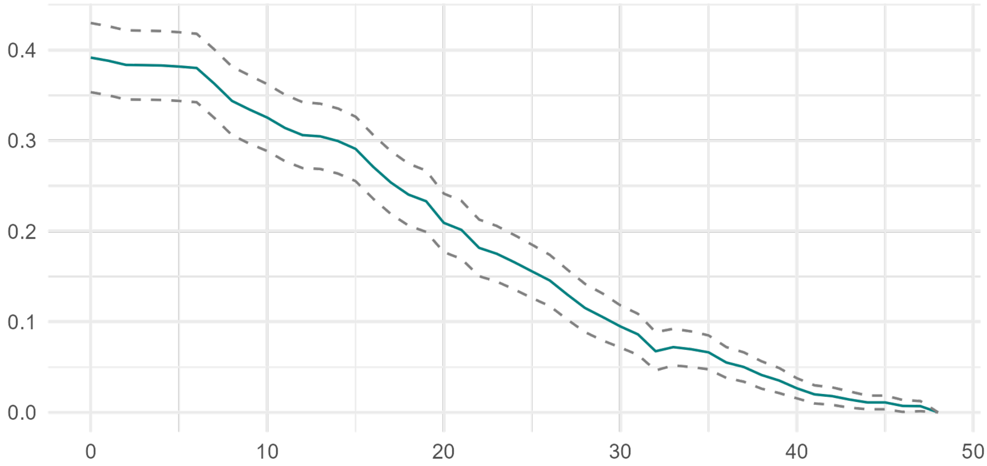

Figure 6: The difference of the conditional CDFs gives the unstandardized weighting function of the CC-ACR parameter in Equation (6). Standardizing these values to sum up to one gives the weights on the average causal responses of the combined complier types. 95% confidence intervals are calculated in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

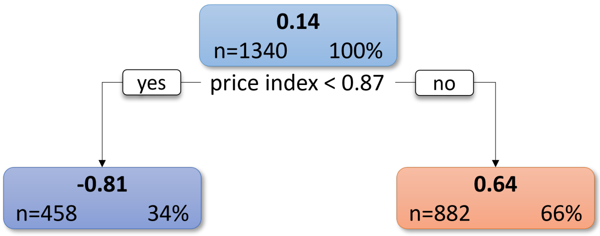

Figure 7: Plot of the shallow Breiman tree that led to the maximum local violation of the LiM assumption. The variable name under the leaf depicts the variable and the value that the leaf subsequently was split on. The top value in each leaf shows a violation of LiM if it is positive and the magnitude of the violation. The bottom row gives the number of observations in the leaf, both in absolute and relative terms.

A violation of LiM is detected when splitting the data based on the food price index. The tree that led to the greatest local violation is visualized in Figure 7. This violation was found at $j = 47$, and comparable violations are observed across all other treatment levels. The estimate $\hat{\tau}_{47}(\text{food price index} \geq 0.87) = 0.64$, resulting from Equation (9), indicates that the greatest violation was observed in the subgroup where the food price index exceeded 0.87. Generally, the price index may be lower in poor regions than in rich regions. Therefore, this particular violation might hint at children from wealthier households violating the LiM assumption as they might defy the capacity instrument, as hypothesized. It should be emphasized that these results are purely indicative.

7 Conclusion

This study set out to gain a better understanding of what can be identified in the setting involving multiple instruments and nonbinary, ordered treatments. The pivotal theoretical contribution of this study is the identification of the average causal response for combined complier populations, the CC-ACR, which offers an attractive alternative to the TSLS approach, for which this study presented novel insights. Unlike TSLS, the CC-ACR allows for imposing a less restrictive monotonicity assumption and presents a more intuitive weighting scheme, making it a valuable tool for obtaining results that can be readily interpreted as the average causal effect for a one-level increase in the treatment for the combined complier population. Moreover, this study went beyond identification and provided practical guidance for estimation, leveraging recent advances in causal machine learning. The empirical application, which explored the causal effect of HC nurseries on a child's height – a widely used proxy for nutritional status – revealed that the program does not yield a significant effect across the different estimation methods. This insight holds substantial policy relevance, as it suggests a need for potential restructuring of the program if its primary aim is to enhance nutritional outcomes.

A further notable theoretical contribution of this study lied in the development of a test for LiM, a fundamental assumption underlying the identification of the CC-ACR. This novel test, which uses causal forests and detects local violations in a data-driven manner, successfully detected a potential violation of LiM. Consequently, it is crucial to interpret the empirical estimates with caution, as the direction of LiM may vary within different subgroups of the population.

Despite the contributions this study makes to the existing literature, it is important to recognize its limitations. These limitations warrant careful consideration as they point toward avenues for future research and potential refinements in methodology. For instance, a major limitation of this study is that only a subsample of the data is used for estimation. A natural progression of this work is to investigate what can be learned from the observations that currently are discarded. Potentially, one could form different subsets of the instruments where LiM holds and then combine the resulting CC-ACRs into one aggregated estimate. A similar approach has been undertaken by Sun and Wüthrich (2022) under the IAM assumption. This approach would allow for maximizing the information used by also considering combined complier types at intermediate instrument values.

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 (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 et al. (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.

References

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

Andresen, M. E. and Huber, M. (2021). Instrument-based Estimation with Binarised Treatments: Issues and Tests for the Exclusion Restriction. *The Econometrics Journal*, 24(3):536–558.

Angrist, J. D., Graddy, K., and 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.

Angrist, J. D. and 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.

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of Causal Effects Using Instrumental Variables. *Journal of the American statistical Association*, 91(434):444–455.

Athey, S., Tibshirani, J., and Wager, S. (2019). Generalized Random Forests. *The Annals of Statistics*, 47(2):1148–1178.

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

Bach, P., Chernozhukov, V., Kurz, M. S., and Spindler, M. (2021). DoubleML – An Object-Oriented Implementation of Double Machine Learning in R. arXiv:2103.09603 [stat.ML].

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

Bhuller, M. and Sigstad, H. (2022). 2SLS with Multiple Treatments. *arXiv preprint arXiv:2205.07836*.

Bhutta, Z. A., Ahmed, T., Black, R. E., Cousens, S., Dewey, K., Giugliani, E., Haider, B. A., Kirkwood, B., Morris, S. S., Sachdev, H., et al. (2008). What Works? Interventions for Maternal and Child Undernutrition and Survival. *The lancet*, 371(9610):417–440.

Blandhol, C., Bonney, J., Mogstad, M., and Torgovitsky, A. (2022). *When is TSLS Actually LATE?* (No. w29709). National Bureau of Economic Research.

Breiman, L. (2001). Random Forests. *Machine Learning*, 45:5–32.

Carneiro, P., Heckman, J. J., and Vytlacil, E. J. (2011). Estimating Marginal Returns to Education. *American Economic Review*, 101(6):2754–2781.

Carr, T. and Kitagawa, T. (2021). Testing Instrument Validity with Covariates. *arXiv preprint arXiv:2112.08092*.

Chen, T. and Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/Debiased Machine Learning for Treatment and Structural Parameters.

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

Farbmacher, H., Guber, R., and Klaassen, S. (2022). Instrument Validity Tests with Causal Forests. *Journal of Business & Economic Statistics*, 40(2):605–614.

Frandsen, B., Lefgren, L., and Leslie, E. (2023). Judging Judge Fixed Effects. *American Economic Review*, 113(1):253–277.

Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *Annals of Statistics*, pages 1189–1232.

Frölich, M. (2007). Nonparametric IV Estimation of Local Average Treatment Effects with Covariates. *Journal of Econometrics*, 139(1):35–75.

González Ramírez, J. L. and Durán, I. M. (2012). Evaluation for Improving: The Impact Evaluation of the Program Hogares Comunitarios de Bienestar of ICBF. *Desarrollo y Sociedad*, (69):187–234.

Heckman, J. J. and Pinto, R. (2018). Unordered Monotonicity. *Econometrica*, 86(1):1–35.

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

Hotelling, H. (1933). Analysis of a Complex of Statistical Variables into Principal Components. *Journal of Educational Psychology*, 24(6):417.

Huber, M. and Kueck, J. (2022). Testing the Identification of Causal Effects in Observational Data. *arXiv preprint arXiv:2203.15890*.

Huber, M., Laffers, L., and 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.

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

Imbens, G. and Angrist, J. (1994). Identification and Estimation of Local Average Treatment Effects. *Econometrica*, 62(2):467–476.

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

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

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

Lee, S. and Salanié, B. (2018). Identifying Effects of Multivalued Treatments. *Econometrica*, 86(6):1939–1963.

Martins, V. J., Toledo Florêncio, T. M., Grillo, L. P., Franco, M. d. C. P., Martins, P. A., Clemente, A. P. G., Santos, C. D., Vieira, M. d. F. A., and Sawaya, A. L. (2011). Long-Lasting Effects of Undernutrition. *International Journal of Environmental Research and Public Health*, 8(6):1817–1846.

Mogstad, M., Torgovitsky, A., and Walters, C. R. (2021). The Causal Interpretation of Two-Stage Least Squares with Multiple Instrumental Variables. *American Economic Review*, 111(11):3663–98.

Mountjoy, J. (2022). Community Colleges and Upward Mobility. *American Economic Review*, 112(8):2580–2630.

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

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

Prado, E. L. and Dewey, K. G. (2014). Nutrition and Brain Development in Early Life. *Nutrition Reviews*, 72(4):267–284.

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.

Robins, J. M., Rotnitzky, A., and 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.

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

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

Sun, Z. (2020). Instrument Validity for Heterogeneous Causal Effects. *arXiv preprint arXiv:2009.01995.*

Sun, Z. and Wüthrich, K. (2022). Pairwise Valid Instruments. *arXiv preprint arXiv:2203.08050.*

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

van 't Hoff, N., Lewbel, A., and Mellace, G. (2023). Limited Monotonicity and the Combined Compliers LATE. Boston College Working Papers in Economics 1059.

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

Wager, S., Athey, S., and Tibshirani, J. (2019). Generalized Random Forests. `https://github.com/grf-labs/grf`. Software version 2.2.1.

Appendices

A Complete table with response types

Table 6: 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	✓	✓	
	$c_{0,1,1,2}$	0	1	1	2	✓	✓	✓

Continued on next page

Table 6 – continued from previous page

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

Continued on next page

Table 6 – continued from previous page

Combined type	Type	D^{00}	D^{10}	D^{01}	D^{00}	LiM	PM	IAM
	$d_{2,0,2,0}$	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 $\mathscr{Z}$ 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:

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

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 $\mathscr{T}$ the set of all response types t, and $\sum_{t \in \mathscr{T}} P(T = t) = 1$. 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:

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

B.4 Proof of alternative representation of the TSLS estimand

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

$$\beta_{TSLS} \equiv \sum_{m=1}^{M} \delta_{m,m-1} \cdot \omega_m \cdot \beta_{m,m-1}, \tag{10}$$

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

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

This can be seen as follows:

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

$$= E\Bigg(Y^{j-1} \cdot (I(D^{z_m} \geq 0) - I(D^{z_m} \geq 1) - I(D^{z_{m-1}} \geq j-1) - I(D^{z_{m-1}} \geq 1))$$

$$+ \sum_{j=1}^{J} Y^j \cdot (I(D^{z_m} \geq j) - I(D^{z_m} \geq j+1) - I(D^{z_{m-1}} \geq j) - I(D^{z_{m-1}} \geq j+1))\Bigg)$$

$$= E\Bigg(Y^{j-1} \cdot (I(D^{z_m} \geq j-1) - I(D^{z_{m-1}} \geq j-1))$$

$$+ \sum_{j=1}^{J} (Y^j - Y^{j-1}) \cdot (I(D^{z_m} \geq j) - I(D^{z_{m-1}} \geq j))\Bigg)$$

$$= E(\sum_{j=1}^{J} (Y^j - Y^{j-1}) \cdot (I(D^{z_m} \geq j) - I(D^{z_{m-1}} \geq j)))$$

$$= \sum_{j=1}^{J} E((Y^j - Y^{j-1}) \cdot (I(D^{z_m} \geq j) - I(D^{z_{m-1}} \geq j)))$$

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

$$= \sum_{j=1}^{J} \Big\{ 1 \cdot P(I(D^{z_m} \geq j) - I(D^{z_{m-1}} \geq j) = 1)$$

$$\cdot E((Y^j - Y^{j-1}) \cdot (I(D^{z_m} \geq j) - I(D^{z_{m-1}} \geq j)) | I(D^{z_m} \geq j) - I(D^{z_{m-1}} \geq j) = 1)$$

$$+ 0 \cdot P(I(D^{z_m} \geq j) - I(D^{z_{m-1}} \geq j) = 0)$$

$$\cdot E((Y^j - Y^{j-1}) \cdot (I(D^{z_m} \geq j) - I(D^{z_{m-1}} \geq j)) | I(D^{z_m} \geq j) - I(D^{z_{m-1}} \geq j) = 0)$$

$$- 1 \cdot P(I(D^{z_m} \geq j) - I(D^{z_{m-1}} \geq j) = -1)$$

$$\cdot E((Y^j - Y^{j-1}) \cdot (I(D^{z_m} \geq j) - I(D^{z_{m-1}} \geq j)) | I(D^{z_m} \geq j) - I(D^{z_{m-1}} \geq j) = -1) \Big\}$$

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

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

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

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

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 (10) 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 et al. (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)):

$$P(D < j|Z_1 = Z_2 = ... = Z_K = 0) - P(D < j|Z_1 = Z_2 = ... = Z_K = 1) \geq 0, \tag{11}$$

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

The inequalities in Condition (11) 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 et al. (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 et al. (1994):

$$\begin{aligned}\hat{\Gamma}_{j,i} \equiv & \hat{\tau}_j^{(-i)}(X_i) \\ & + \frac{\widetilde{Z}_i - \hat{e}^{(-i)}(X_i)}{\hat{e}^{(-i)}(X_i)\left(1 - \hat{e}^{(-i)}(X_i)\right)} \times \left(Q_{j,i} - \hat{\mu}_j^{(-i)}(X_i) - (\widetilde{Z}_i - \hat{e}^{(-i)}(X_i))\hat{\tau}_j^{(-i)}(X_i)\right).\end{aligned} \tag{12}$$

$\hat{\tau}_j(X_i)$, $\hat{e}(X_i)$, 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 Π_j^{SA} and Π_j^{SB}. 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 et al. (2022) using the results of Chernozhukov et al. (2018).

C.3 Pseudo code of the LiM testing procedure

I build upon the procedure and code of Farbmacher et al. (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 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 (12) 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 $T_{\text{vec}}$.
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
```

D Covariate table

Table 7: Overview of the controls included in the analysis. Variable descriptions are borrowed from Attanasio et al. (2013).

Variable	Description	Mean for $z_0 = (0,0,0)$	Mean for $z_M = (1,1,1)$	p-value for the difference
Female	1 if child is female, 0 if child is male	0.517	0.483	0.156
Age mother	Mother's age in years divided by 100	0.322	0.324	0.488
Age head	Household head s age in years divided by 100	0.389	0.404	0.004
Mother's height	Mother's height in meters	1.544	1.546	0.602
Order	Order of child in the household	5.940	5.988	0.791
Educ. head below secondary	1 if mother completed primary education but did not complete secondary education, 0 otherwise	0.386	0.313	0.001
Educ. head secondary	1 if mother completed secondary education, 0 otherwise	0.034	0.097	0.000
Educ. mother below secondary	1 if household head completed primary education but did not complete secondary education, 0 otherwise	0.307	0.284	0.292
Educ. mother secondary	1 if household head completed secondary education, 0 otherwise	0.025	0.098	0.000
Centre town	1 if household lives in the main part of the town, 0 otherwise	0.427	0.673	0.000
Distance health center (mins)	Distance in minutes to the nearest health care provider, divided by 100	0.407	0.240	0.000
Distance school center (mins)	Distance in minutes to nearest school, divided by 100	0.176	0.076	0.000
Distance town hall (mins)	Distance in minutes to the town hall, divided by 100	0.470	0.356	0.000
Hospital	1 if there is a hospital in the town, 0 otherwise	0.735	0.613	0.000
Health insurance (prop. In town)	Proportion of children with formal health insurance in the municipality	0.581	0.672	0.000
Number of children	Number of children aged 2 to 6 years old in the town, divided by 10,000	0.318	0.168	0.000
Wage female rural	Urban female daily wage in pesos as indicated by the town major divided by 10,000 in Colombian pesos (December 2003)	1.023	0.881	0.000
Wage female urban	Rural female daily wage in pesos as indicated by the town major divided by 10,000 in Colombian pesos (December 2003)	0.928	0.826	0.000
Price index	Food price index	0.920	0.921	0.876
Region 2	Region 2	0.289	0.031	0.000
Region 3	Region 3	0.452	0.026	0.000
Region 4	Region 4	0.068	0.200	0.000
Second wave	Second wave	0.395	0.490	0.000
Third wave	Third wave	0.200	0.132	0.000

E Details on the model specifications

E.1 Specifications of the methods employed in Section 6.2

E.1.1 Specifications of the GRF

In a nutshell, GRFs (Athey et al., 2019) provide a GMM framework with forest-constructed weights. Estimation is performed using the package as provided by Wager et al. (2019). Following the approach of Attanasio et al. (2013), clustering is conducted at the municipality level. The forest comprises the default number of 2,000 trees, with the remaining parameters obtained via cross-validation. Table 8 provides summary statistics for all tuned parameters based on 100 bootstrap runs.

Table 8: Summary statistics for the tuned GRF parameters over the 100 bootstrap repetitions for the results presented in Section 6.2.

Outcome	Parameter	Mean	Std.	Min.	Max.
Y	Sample fraction	0.31	0.16	0.05	0.50
	Mtry	16.33	7.65	1.00	24.00
	Minimum node size	20.00	31.52	1.00	175.00
	Honesty fraction	0.61	0.10	0.50	0.80
	Honesty prune leaves	0.67	0.47	0.00	1.00
	Alpha	0.09	0.06	0.00	0.25
	Imbalance penalty	0.56	0.79	0.00	4.47
D	Sample fraction	0.36	0.15	0.05	0.50
	Mtry	16.00	8.23	1.00	24.00
	Minimum node size	17.58	26.82	1.00	156.00
	Honesty fraction	0.59	0.09	0.50	0.80
	Honesty prune leaves	0.69	0.46	0.00	1.00
	Alpha	0.10	0.07	0.00	0.25
	Imbalance penalty	0.57	0.75	0.00	4.40

E.1.2 Specifications of the random forest within the DML framework

The core idea of the DML approach is to orthogonalize both the treatment and the outcome variable, and rely on sample-splitting for debiasing. In this study, I utilize the DML method (Chernozhukov et al., 2018), assuming a partially linear model, in combination with the random forest algorithm (Breiman, 2001) to estimate both the numerator and denominator components of the CC-ACR parameter. The analyses are conducted using the DoubleML package as provided by Bach et al. (2021). The results presented in Section 6.2 are obtained through a sample splitting procedure with a single repetition and a 5-fold cross-fitting approach. Default parameters for the random forest are as follows:

- Alpha: 0.5
- Minimum node size: 5
- Mtry: 5
- Sample fraction: 1
- Number of trees: 500

E.1.3 Specifications of the instrumental forest

I implement the instrumental forest of the GRF framework (Athey et al., 2019), using the package as provided by Wager et al. (2019). Following the approach of Attanasio et al. (2013), clustering is conducted at the municipality level. The parameters are set to their default values:

- Number of trees: 2,000
- Sample fraction: 0.5
- Mtry: 24
- Minimum node size: 5
- Honesty fraction: 0.5
- Honesty prune leaves: 1
- Alpha: 0.05
- Imbalance penalty: 0

E.1.4 Specification of the GRF when the treatment is continuous

The same approach as outlined in Appendix E.1.1 can be taken when the treatment is continuous. The summary statistics for the tuned parameters over the 100 bootstrap repetitions are presented in Table 9.

Table 9: Summary statistics for the tuned GRF parameters over the 100 bootstrap repetitions for the continuous treatment results in Section 6.2.

Outcome	Parameter	Mean	Std.	Min.	Max.
Y	Sample fraction	0.24	0.15	0.05	0.50
	Mtry	12.75	7.08	1.00	24.00
	Minimum node size	37.27	39.83	1.00	175.00
	Honesty fraction	0.64	0.09	0.50	0.80
	Honesty prune leaves	0.54	0.50	0.00	1.00
	Alpha	0.10	0.07	0.00	0.25
	Imbalance penalty	0.90	0.94	0.00	4.47
D	Sample fraction	0.31	0.15	0.05	0.50
	Mtry	13.45	7.81	1.00	24.00
	Minimum node size	23.82	34.42	1.00	163.00
	Honesty fraction	0.62	0.09	0.50	0.80
	Honesty prune leaves	0.61	0.49	0.00	1.00
	Alpha	0.11	0.06	0.01	0.25
	Imbalance penalty	0.72	0.98	0.00	6.97

F Identification of a causal effect for unordered treatments

While the present study focuses on ordered treatments, it is straightforward to extend the results to unordered treatments. Recently, there have been important advances in identifying causal effects of unordered treatments. Heckman and Pinto (2018) introduce unordered monotonicity, an assumption similar to IAM. They investigate what counterfactuals can be identified when discrete instruments are available. Note that, in the case of unordered treatments, at least one instrument is required for every treatment value. Mountjoy (2022) builds upon this work and introduces unordered partial monotonicity (UPM). He shows that, when multiple continuous instruments are available, causal effects along multiple treatment margins are identified. In a similar fashion, I generalize LiM to *unordered limited monotonicity* (ULiM). Denote the treatment random variable with T and its support with $\text{supp}(T) = \{t_1, ..., t_{N_t}\}$, where N_t denotes the number of distinct treatments.

Unordered limited monotonicity (ULiM):
For some treatment $t \in \text{supp}(T)$, it holds that

$$I(T_i^{1...1...1} = t) \geq I(T_i^{0...0...0} = t) \text{ or } I(T_i^{1...1...1} = t) \leq I(T_i^{0...0...0} = t), \ \forall i.$$

This means that no units opt out of treatment t when all instrument values are switched from zero to one. This assumption might hold for different treatments for different subsets of the instruments. Similar to the testing procedure for LiM described in the present paper, one can extend Sun's (2020) results to the causal forest testing framework to detect potential violations of ULiM. It can further be shown that $E(Y^t|t\text{-compliers})$ is identified under ULiM. This is the mean counterfactual outcome for those individuals who do change their treatment status in response to a particular change in the instrument values, the so-called t-compliers:

$$\begin{aligned}
&E(Y^t|t\text{-compliers})\\
&= \frac{E(Y^t \cdot I(T=t)|Z_1=1, Z_2=1, ..., Z_K=1) - E(Y^t \cdot I(T=t)|Z_1=0, Z_2=0, ..., Z_K=0)}{P(T=t|Z_1=1, Z_2=1, ..., Z_K=1) - P(T=t|Z_1=0, Z_2=0, ..., Z_K=0)}\\
&= \sum_{i=1}^{K-1} E\left(Y^t|s \in \sum_t(i)\right) \cdot \frac{P(s \in \sum_t(i))}{P(S \in t\text{-compliers})},
\end{aligned}$$

where the response types are denoted $s \in \text{supp}(S)$ and $\sum_t(i)$ indicates the set of response types for which the treatment t occurs precisely i times (refer to Heckman and Pinto (2018) for more details). Subsequently, the local average treatment effect can be identified for the t-compliers, $E(Y^t - Y^{t'}|t\text{-compliers})$ with t' all treatments that are not t, providing the effect of t compared to the next best option in t'.