Limited Monotonicity and the Combined Compliers LATE*

Nadja van 't Hoff[1], Arthur Lewbel[2], and Giovanni Mellace[1]

[1] *University of Southern Denmark*

[2] *Boston College*

May 24, 2023

Abstract

We consider estimating a local average treatment effect given an endogenous binary treatment and two or more valid binary instruments. We propose a novel limited monotonicity assumption that is generally weaker than alternative monotonicity assumptions considered in the literature, and allows for a great deal of choice heterogeneity. Using this limited monotonicity, we define and identify the Combined Complier Local Average Treatment Effect (CC-LATE), which is arguably a more policy relevant parameter than the weighted average of LATEs identified by Two Stage Least Squares. We apply our results to estimate the effect of learning one's HIV status on protective behaviors.

Keywords: Instrumental variable, Local Average Treatment Effect, monotonicity, multiple instruments.

JEL classification: C14, C21, C26.

*Corresponding author Giovanni Mellace (giome@sam.sdu.dk). This project was made possible through generous funding by Independent Research Fund Denmark (90380031B). We have benefited from discussions with Tymon Sloczyński, Phillip Heiler, Jonathan Roth, and participants at the Nordic Econometric Meeting 2022, the European Winter Meeting of the Econometric Society 2022.

1 Introduction

Instrumental variables are commonly used to address endogeneity issues in the treatment variable. Endogeneity arises when the treatment is not randomly assigned and individuals self-select into treatment based on observed and unobserved characteristics. In many settings it is more realistic that treatment effects vary across individuals based on both observed and unobserved factors. When treatment effects are heterogeneous and multiple valid (unconfounded) instruments are available, each instrument separately identifies the effect for the individuals whose treatment status changes in response to the instrument. The treatment effect in the subgroup of these compliers is referred to as the local average treatment effect (LATE). The usual practice for combining instruments is two-stage least-squares (TSLS). Imbens and Angrist (1994) show that TSLS identifies a weighted average of the the instrument-pair LATEs in the case of multiple binary instruments. They introduce the monotonicity assumption which ensures that individuals respond to a change in the instrument in a monotone way, meaning that two-way flows in response to a change in the instrument are ruled out. We follow Mogstad et al. (2021) in referring to this monotonicity assumption as IAM (Imbens and Angrist Monotonicity).

Despite being common practice for combining multiple instruments, TSLS has several shortcomings. Assuming IAM is equivalent to assuming choice homogeneity. While treatment effects are commonly allowed to be heterogeneous, choices are not. This asymmetry is pointed out by Heckman et al. (2006). Mogstad et al. (2021) explain why choice homogeneity can be undesirable to assume, using the returns to education example with tuition subsidy and college proximity as instruments. The choice restrictions imposed by IAM imply that all individuals either favor tuition subsidy or college proximity. IAM does not allow for one individual to favor the first incentive over the latter while another individual favors the latter incentive over the first. It hence heavily restricts the response types that are present in the population. Mogstad et al. (2021) relax IAM to the weaker partial monotonicity (PM) assumption that allows for more choice heterogeneity. PM considers a change in a single component of the instrument while holding the values of the other instruments fixed. Put another way, PM implies random coefficients with restricted signs in the selection equation, whereas IAM additionally restricts the magnitude of the coefficients. They further show that the TSLS estimand retains the interpretation of a weighted average of individual LATEs in the case of multiple binary instruments, with the LATEs corresponding to different response groups. However, as will be outlined later in the paper, there are various applications where PM still might be too restrictive. To mention one example: PM does not allow for the plausible scenario where some individuals attend college when it is cheap and far, while other individuals attend college only when it is cheap and close.

Another issue with using TSLS with multiple instruments is whether the identified parameter

is interesting. For instance, Heckman et al. (2006) doubt whether a weighted average of LATEs is of economic interest. The weights in the weighted average of LATEs given by the TSLS estimand are counter-intuitive. For instance, the weights may well be negative. Heckman et al. (2006) point out that monotonicity does not necessarily imply positive weights and vice versa. The interpretation of the TSLS estimand is further complicated when the weighted average of LATEs estimated by TSLS includes the LATEs of defier types.

The purpose of the present paper is to address the shortcomings of linear IV models with multiple instruments by providing a more credible monotonicity assumption and an estimand with a more intuitive interpretation than the weighted average of LATEs identified by TSLS. The proposed monotonicity assumption is referred to as *limited monotonicity* (LiM). This assumption only imposes that the treatment status of all units when exposed to all instruments simultaneously is greater than or equal to the treatment status when exposed to none of the instruments. Since it does not impose any restrictions on choice behavior when units are exposed to a subset of the instruments, LiM allows for rich choice heterogeneity. LiM requires fewer choice restrictions than PM, allowing for many more response types in the population.

Under LiM, we show that a parameter called the Combined Compliers Local Average Treatment Effect (CC-LATE) is identified, and provide a very simple consistent estimator. The CC-LATE is defined as the ATE for those individuals who do not take up treatment in the absence of the instruments but whose treatment status changes when exposed to all instruments combined. We refer to this set of individuals as "combined compliers". The combined compliers represent the target population that is pushed towards the policy by the instruments. This is especially of interest when the instruments can be manipulated by the policy-maker. Moreover, it is likely (but not guaranteed) that the larger the group of compliers, the closer the CC-LATE is to the ATE, especially when the treatment effect does not vary much across the population. This is because the larger is the set of compliers, the closer the set of compliers becomes to representing the population as a whole. Since the CC-LATE gives the effect amongst those individuals that are affected by the instruments, it refers to a large complier population.

The CC-LATE is a much more interesting and broadly applicable parameter for a policy-maker than the TSLS estimand for two reasons. Firstly, the interpretation of the CC-LATE is straightforward and intuitive, as opposed to the interpretation of the TSLS estimand. The CC-LATE can be interpreted as a weighted average over the combined complier LATEs with the weights equalling the corresponding complier shares. Thus the weights are non-negative by construction and have an intuitive interpretation. Secondly, the CC-LATE is still identified in the presence of a variety of defier types. This is an attractive property of the CC-LATE since the number of potential defier types grows rapidly with the number of available instruments.

We illustrate these results by estimating the effect of learning one's HIV status on protective

behavior like the purchase of condoms. Thornton (2008) investigates the effect of knowing one's HIV status on the purchase of contraceptives in rural Malawi, countering selection issues by instrumenting with a financial incentive offered in the form of cash and with the distance to the recommended HIV center. We argue that LiM might be more plausible than PM in this application. We find that the CC-LATE estimates provide more evidence for protective behavior after learning one's HIV statis than the TSLS estimates. Differences between the estimates might be due to negative TSLS weights and/or a violation of PM.

Our work is most closely related to that of Mogstad et al. (2021), Frölich (2007), and Goff (2020). Mogstad et al. (2021) introduce PM and show that the TSLS estimand retains the interpretation of a weighted average of LATEs under this assumption. For reasons discussed above, LiM is generally less restrictive than PM and the CC-LATE is a more intiutive parameter than the weighted average of LATEs that TSLS identifies. Frölich (2007) considers identification with multiple instrumental variables. His estimand is similar to ours, but it differs in that it relies on IAM. It also differs in terms of interpretation, i.e., Frölich (2007)'s proposed estimand gives the effect for the largest group of compliers, whereas the CC-LATE considers the entire complier population and hence includes at least as many individuals. Goff (2020) introduces vector monotonicity (VM) which is a special form of PM. VM is strictly stronger than LiM and heavily restricts the response types by ruling out the presence of many defier types. Under this assumption Goff (2020) shows that the "all compliers" LATE (ACL) is identified. In the setting with two binary instruments, the combined complier population of the CC-LATE is similar to Goff (2020)'s all complier population, and the interpretation of the ACL and the CC-LATE coincides. When more than two instruments are available, the CC-LATE gives the ATE for a larger complier population which is generally more desirable.

Other literature has focused on relaxing the monotonicity assumption in the setting with a binary treatment and a single binary instrument (Słoczyński, 2020; Kolesár, 2013; Small et al., 2017; De Chaisemartin, 2017; Dahl et al., 2017), or on relaxing or omitting monotonicity in the case of unordered treatments (Kirkeboen et al., 2016; Hull, 2018; Salanié and Lee, 2018; Heckman and Pinto, 2018). In the multiple instruments setting, Huntington-Klein (2020) derives identification of the Super-Local Average Treatment Effect under a condition where monotonicity is imposed within subgroups of the data. Mogstad et al. (2020) show that each instrument has its own selection equation under partial monotonicity and use mutual consistency of these equations to obtain information about (instrument-invariant) parameters. There is some literature looking to estimate the treatment effect beyond the complier population through extrapolation. For instance, Mogstad and Torgovitsky (2018) extrapolate the support of a single LATE to include observations other than compliers and provide bounds. Mogstad et al. (2018) extrapolate the LATE to a population with lower willingness to pay for treatment.

The remainder of this paper is organized as follows: The study will begin by introducing the limited monotonicity assumption and the Combined Compliers LATE for the setting with two binary instruments in Section 2. This is followed by a comparison of LiM to other versions of the monotonicity assumption in Section 3. Section 4 extends the CC-LATE to the setting with multiple binary instruments and Section 5 then compares the CC-LATE estimand to other estimands. Section 6 provides an empirical application to the impact of learning HIV status on contraceptive use as considered by Thornton (2008). Finally, Section 7 provides a discussion and Section 8 adds some extensions.

2 Combined Compliers LATE

2.1 Definitions and baseline assumptions

Consider the standard Imbens and Angrist (1994) LATE framework, including an outcome Y and a binary treatment D. Assume we have k binary instruments $Z_1, Z_2, ..., Z_k$. Denote by $D_i^{z_1 z_2 ... z_k} \in \{0,1\}$ the potential treatment states, and by $Y_i^{d, z_1 z_2 ... z_k}$ the potential outcomes (see for instance Rubin, 1974), assuming that the instruments satisfy the exclusion restriction, i.e., they do not directly affect Y_i^d, and are independent of the potential treatments and outcomes. This ensures that the instruments are as good as randomly assigned. Formally this is given by Assumption 1[1].

Assumption 1: Random assignment and exclusion

$Z_j \perp (D^{z_1 z_2 ... z_k}, Y^d) \quad \forall z_1, z_2, ..., z_k, d \in \{0,1\}, j \in \{1,2,...,k\}.$

We make the following two additional assumptions, which are standard for LATE estimation: The stable unit treatment value assumption (SUTVA) and the instrument relevance assumption. SUTVA ensures that the treatment assigned to any individual does not affect the potential outcomes of any other individual, that the individuals do not potentially have access to different version of the treatment, and that there is no measurement error. The relevance assumption ensures that compliers exist.

Assumption 2: SUTVA

$Y = Y^d$ if $D = d$, and $D = D^{z_1 z_2 ... z_k}$ if $Z_1 = z_1$, $Z_2 = z_2$, ..., and $Z_k = z_k$.

[1]Assumption 1 can be replaced by mean independence when mean effects are of interest, as is the case in our setting. However, in many settings, making the stronger assumption of independence is as realistic as imposing mean independence.

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 $P(D^{1...1...1} = 1) \neq P(D^{0...0...0} = 1)$.

These three assumptions alone do not guarantee identification of an interpretable parameter of a causal effect. With one binary instrument, the monotonicity assumption (that the treatment indicator is always greater than or equal to the instrument) rules out defiers. With multiple binary instruments, we impose the monotonicity condition that individuals are as least as likely to be treated if all the instruments are switched on than when all the instruments are switched off. In terms of potential treatments, this gives Assumption 4. We refer to this assumption as *limited* monotonicity, since it only imposes a constraint on $P(D^{1...1...1} \geq D^{0...0...0})$. In Section 3, we compare LiM to the monotonicity assumptions proposed by Imbens and Angrist (1994) and Mogstad et al. (2021), and show that LiM is generally weaker.

Assumption 4: Limited monotonicity (LiM)

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

Throughout we assume that the instruments are defined such that positive LiM holds, i.e, $P(D^{1...1...1} \geq D^{0...0...0}) = 1$. This only requires to define all instruments such that they have a positive first stage.

2.2 Two binary instrument setting

We first demonstrate our results for the two binary instrument setting. These results are generalizable to an arbitrary number of binary instruments as shown in Section 4.

2.2.1 Principal strata and types

With one binary instrument, Imbens and Angrist (1994) (see also Angrist et al., 1996) define four types of individuals: compliers, always-takers, never-takers, and defiers. These types are defined by the values of their potential treatments. With two binary instruments there are sixteen possible types of individuals, as listed in Table 1. Similar to the setting with one binary instrument, the never-takers (nt) never take up treatment and the always-takers (at) always take up treatment, independent of the instrument values. We follow Mogstad et al. (2021) in labeling some of the other response types: The eager compliers (ec), the reluctant compliers (rc), the first instrument compliers ($1c$), and the second instrument compliers ($2c$). These compliers respond to either one of the instruments or a combination thereof. We define combined compliers as the set $cc \equiv \{ec, rc, 1c, 2c\}$, so combined compliers are any of these four complier types.

There are different defier types with two binary instruments. Second instrument defiers ($2d$) respond more strongly to the first instrument, since $D = 1$ when $Z_1 = 1$ ($D^{11} = 1$ and $D^{10} = 1$), but they are defiers with respect to the second instrument as soon as $Z_1 = 0$ ($D^{01} = 0$ and $D^{00} = 1$). Similar reasoning can be followed for the first instrument defiers ($1d$). Eager defiers (ed) only take up treatment when either both instruments are switched on ($D^{11} = 1$) or when both instruments are switched off ($D^{00} = 1$), but not when a single instrument is switched on ($D^{10} = 0$ and $D^{01} = 0$). Reluctant defiers (rd) do not take up treatment when either both instruments are switched on ($D^{11} = 0$) or when both instruments are switched off ($D^{00} = 0$), but they do take up treatment when a single instrument is switched on ($D^{10} = 1$ and $D^{01} = 1$). Finally, there are six other defier types ($d1$, $d2$, $d3$, $d4$, $d5$, and $d6$).

Note that unlike the case with a single binary instrument, monotonicity with multiple instruments means that there are more defier types than complier types. This is due to the existence of defiers with respect to either instrument. When only one of the instruments is observed such that we are in the single binary instrument setting, the type corresponding to the observed instrument might differ depending on the value of the unobserved instrument (see Table 1). For instance, consider an eager defier (ed). If only instrument Z_1 were observed, this individual would be a complier when the value of the unobserved instrument Z_2 equals one ($Z_2 = 1$). The same individual would be a defier with respect to Z_1 when $Z_2 = 0$. Adding instruments hence allows for more flexibility in defier types and provides more detailed response types.

In the two instrument setting, LiM reduces to the following assumption[2]:

Limited monotonicity (LiM) in the two instrument setting

$P(D^{11} \geq D^{00}) = 1$.

LiM allows for 12 out of the 16 initial response types (see Table 1). It rules out four defier types, as shown in Table 1 ($d3$, $d4$, $d5$, and $d6$). These are the defier types that would take up treatment when all instruments are switched off ($D^{11} = 0$), but that would not take up treatment when all instruments are switched on ($D^{00} = 1$). These response types never classify as a complier when only one of the instruments is observed. More specifically, receiving a second instrument never pushes these individuals towards compliance.

[2]Vytlacil's equivalence result (Vytlacil, 2002) connects the LATE assumptions to selection models. Monotonicity assumptions place restrictions on choice behavior. Suppose that we have the following selection equation:

$$D_i(z_1, z_2) = \mathbb{1}[\beta_{0i} + \beta_{1i} z_1 + \beta_{2i} z_2 + \beta_{3i} z_1 z_2 \geq 0].$$

LiM only imposes that either $\beta_{1i} + \beta_{2i} + \beta_{3i} \geq 0$ or $\beta_{1i} + \beta_{2i} + \beta_{3i} \leq 0$. It neither imposes restrictions on the signs and magnitudes of the coefficients nor on direct comparisons between the coefficients. β_{0i}, β_{1i}, β_{2i}, and β_{3i} are allowed to vary with i, allowing for rich choice heterogeneity.

2.2.2 The CC-LATE

Our parameter of interest, denoted by β, is the Combined Compliers Local Average Treatment Effect (CC-LATE), defined as $E(Y^1 - Y^0|T \in cc)$ where T denotes type and the combined compliers are the set $cc \equiv \{ec, rc, 1c, 2c\}$ for the case of two instruments. In this case the CC-LATE corresponds to the ATE for those individuals who are a complier with respect to at least one of the instruments or a combination thereof, whilst not defying the other instrument. In general, the combined compliers are individuals that become compliers when all the instruments are turned on. This implies that the CC-LATE is robust to the presence of all defier types that can be pushed towards compliance with respect to one of the instruments by changing the value of another instrument (see Table 1).

Theorem 1 gives our main result for the setting with two binary instruments.

Theorem 1: Let Assumptions 1, 2, 3, and 4 hold. Then we define the Combined Compliers Local Average Treatment Effect (CC-LATE) as

$$\beta = \frac{E\left(Y \mid Z_1 = 1, Z_2 = 1\right) - E\left(Y \mid Z_1 = 0, Z_2 = 0\right)}{E\left(D \mid Z_1 = 1, Z_2 = 1\right) - E\left(D \mid Z_1 = 0, Z_2 = 0\right)} = E(Y^1 - Y^0|T \in cc),$$

where T denotes type and the combined compliers are the set $cc \equiv \{ec, rc, 1c, 2c\}$.

Proof in Appendix A.

2.2.3 Estimation and inference

A first alternative is to estimate the CC-LATE by applying TSLS in the sample where all instrument values are either equal to zero or equal to one, using $\widetilde{Z} = Z_1 = Z_2 = 1$ (and $\widetilde{Z} = Z_1 = Z_2 = 0$) as the sole instrument. Running TSLS in this subsample gives $\hat{\beta} = (D^T P_{\widetilde{Z}} D)^{-1} D^T P_{\widetilde{Z}} Y$ with $P_{\widetilde{Z}} = \widetilde{Z}(\widetilde{Z}^T \widetilde{Z})^{-1} \widetilde{Z}^T$, which reduces to $\hat{\beta} = (\widetilde{Z}' D)^{-1} \widetilde{Z}' Y$ in the just-identified case. Denote the subsample averages of Y and D when $z_1 = 0$ and $z_2 = 0$ by $\bar{Y}_{00}$ and $\bar{D}_{00}$, and as $\bar{Y}_{11}$, and $\bar{D}_{11}$ when $z_1 = 1$ and $z_2 = 1$. Then $\hat{\beta} = \frac{\bar{Y}_{11} - \bar{Y}_{00}}{\bar{D}_{11} - \bar{D}_{00}}$, as shown in Appendix B. An alternative representation of this estimator using two OLS regressions as well as method of moments (MM) estimation are provided in Appendix C. Standard MM estimation packages can be used to automatically generate consistent estimates and standard errors. It is also possible to estimate the CC-LATE nonparametrically by replacing the expectations of the estimand with sample averages.

Table 1: Principal strata and the definition of the response types in case of two binary instruments and one binary treatment.

Type (T)	D^{11}	D^{10}	D^{01}	D^{00}	Type w.r.t. Z_1 when $Z_2=0$	Type w.r.t. Z_1 when $Z_2=1$	Type w.r.t. Z_2 when $Z_1=0$	Type w.r.t. Z_2 when $Z_1=1$	Notion	LiM	PM	IAM
at	1	1	1	1	always-taker	always-taker	always-taker	always-taker	Always-taker	✓	✓	✓
ec	1	1	1	0	complier	always-taker	complier	always-taker	Eager complier	✓	✓	✓
rc	1	0	0	0	never-taker	complier	never-taker	complier	Reluctant complier	✓	✓	✓
$1c$	1	1	0	0	complier	complier	never-taker	always-taker	First instrument complier	✓	✓	✓
$2c$	1	0	1	0	never-taker	always-taker	complier	complier	Second instrument complier	✓	✓	
$1d$	1	0	1	1	defier	always-taker	always-taker	complier	First instrument defier	✓		
$2d$	1	1	0	1	always-taker	complier	defier	always-taker	Second instrument defier	✓		
ed	1	0	0	1	defier	complier	defier	complier	Eager defier	✓		
rd	0	1	1	0	complier	defier	complier	defier	Reluctant defier	✓		
$d1$	0	1	0	0	complier	never-taker	never-taker	defier	Defier type 1	✓		
$d2$	0	0	1	0	never-taker	defier	complier	never-taker	Defier type 2	✓		
$d3$	0	1	1	1	always-taker	defier	always-taker	defier	Defier type 3			
$d4$	0	1	0	1	always-taker	never-taker	defier	defier	Defier type 4			
$d5$	0	0	1	1	defier	defier	always-taker	never-taker	Defier type 5			
$d6$	0	0	0	1	defier	never-taker	defier	never-taker	Defier type 6			
nt	0	0	0	0	never-taker	never-taker	never-taker	never-taker	Never-taker	✓	✓	✓

✓demonstrates the types allowed for under the respective forms of the monotonicity assumption.

Response types under partial monotonicity underlie the choice restrictions as defined in Equation (1). These are equivalent to Table 3 of Mogstad et al. (2021).

Response types under traditional monotonicity are as introduced by Imbens and Angrist (1994) when all individuals prefer the incentive created by Z_1 over the incentive created by Z_2.

3 Comparison of monotonicity assumptions

3.1 LiM compared to PM and IAM

This section illustrates why LiM is generally more plausible than alternative monotonicity assumptions. The monotonicity assumptions by Imbens and Angrist (1994) and Mogstad et al. (2021) can be formulated as follows [3]:

Imbens and Angrist Monotonocity (IAM) (Imbens and Angrist, 1994)

$P(D^{i...j...k} \geq D^{p...q...r}) = 1$ 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) (Mogstad et al., 2021)

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

Obviously, all three assumptions are equivalent in the case of one binary instrument where they reduce to either $P(D^1 \geq D^0) = 1$ or $P(D^1 \leq D^0) = 1$. When there are two or more instruments, LiM is strictly weaker than IAM. To illustrate this, consider the setting with two binary instruments $Z_1 \in \{0,1\}$ and $Z_2 \in \{0,1\}$ with support $\mathcal{Z} = \{(0,0),(0,1),(1,0),(1,1)\}$. Since there are four different combinations of the instrument values, there are $\binom{4}{2} = 6$ comparisons of potential treatments, $d \in \{0,1\}$. In other words, there are six selection probabilities $P(D^z \geq D^{z'}) = d$ with $z, z' \in \{(0,0),(0,1),(1,0),(1,1)\}$ and $z \neq z'$, that can be restricted by imposing some sort of monotonicity. IAM restricts all six comparisons. LiM always imposes only one restriction, independent of the number of instruments. To give an example, IAM imposes either $P(D^{10} \geq D^{01}) = 1$ or $P(D^{10} \leq D^{01}) = 1$. This translates to requiring that all individuals favor one instrument over the other instrument. Consequently, it is not possible to have some individuals who have a preference for Z_1 and other individuals who have a preference for Z_2. For instance, if all individuals are restricted to favor Z_1 over Z_2, then the response types as shown in Table 1 remain. LiM allows for richer choice heterogeneity through the co-existence of both first instrument compliers and second instrument compliers. Following the same line of reasoning, LiM is less restrictive than the IAM assumption in settings with more than two binary instruments, as it does not impose any ordering on $P(D^{i...j...k} \geq D^{i...j...k})\ \forall i \neq j \neq k$.

[3] Note that vector monotonicity (VM) as introduced by Goff (2020) is equivalent to PM in some settings, and stronger than PM otherwise. Therefore, it is not considered here.

While IAM restricts all six comparisons of potential treatments for different instrument values in the case of two instruments, PM imposes four restrictions. PM requires each of the probabilities $P(D^{00} \geq D^{10})$, $P(D^{00} \geq D^{01})$, $P(D^{10} \geq D^{11})$, and $P(D^{01} \geq D^{11})$ to be either zero or one. Notice that only one of the possible PM assumptions is going to be consistent with the data. Estimating $E(D^{00})$, $E(D^{10})$, $E(D^{01})$, and $E(D^{11})$ reveals the version that is consistent with the data. With two instruments, PM allows for at most seven different response types to co-exist. When increasing the values of the instruments makes participation weakly more likely, PM imposes the following restrictions:

$$P(D^{10} \geq D^{00}) = 1, P(D^{01} \geq D^{00}) = 1, P(D^{01} \geq D^{11}) = 0, P(D^{10} \geq D^{11}) = 0. \tag{1}$$

The six response types consistent with the ordering in Equation (1) are given in Table 1. These choice restrictions rule out six defier types that LiM allows for. It is worth noting that the signs on the choice restrictions $P(D^{10} \geq D^{00}) = 1$ and $P(D^{10} \geq D^{11}) = 0$ as well as $P(D^{01} \geq D^{00}) = 1$ and $P(D^{01} \geq D^{11}) = 0$ are of opposite direction such that $P(D^{00} \geq D^{11}) = 0$ is imposed. PM and LiM are nested in this case and LiM is strictly weaker. So LiM is strictly weaker when increasing (decreasing) instrument values always increases (decreases) treatment uptake.

When the signs of $P(D^{10} \geq D^{00}) = 1$ and $P(D^{10} \geq D^{11}) = 0$ as well as $P(D^{01} \geq D^{00}) = 1$ and $P(D^{01} \geq D^{11}) = 0$ have the same direction, then no restriction on $P(D^{00} \geq D^{11})$ is imposed and LiM and PM are non-nested. With two binary instruments, there are four possible combinations of choice restrictions in accordance with PM that are non-nested with either positive LiM, $P(D^{00} \leq D^{11}) = 1$, or negative LiM, $P(D^{00} \geq D^{11}) = 1$:

$$P(D^{10} \geq D^{00}) = 1,\ P(D^{01} \geq D^{00}) = 1,\ \text{and}\ P(D^{01} \geq D^{11}) = 1,\ P(D^{10} \geq D^{11}) = 1. \tag{2}$$

$$P(D^{10} \geq D^{00}) = 1,\ P(D^{01} \geq D^{00}) = 0,\ \text{and}\ P(D^{01} \geq D^{11}) = 0,\ P(D^{10} \geq D^{11}) = 1. \tag{3}$$

$$P(D^{10} \geq D^{00}) = 0,\ P(D^{01} \geq D^{00}) = 1,\ \text{and}\ P(D^{01} \geq D^{11}) = 1,\ P(D^{10} \geq D^{11}) = 0. \tag{4}$$

$$P(D^{10} \geq D^{00}) = 0,\ P(D^{01} \geq D^{00}) = 0,\ \text{and}\ P(D^{01} \geq D^{11}) = 0,\ P(D^{10} \geq D^{11}) = 0. \tag{5}$$

The response types that are present under these four different versions of the assumptions are listed in Table 2, together with the response types under positive and negative LiM. Clearly, in all four cases, LiM allows for substantial more choice heterogeneity than PM, allowing for a much larger number of different response types. For each of these four versions of PM, there is only one response type that is included under PM that would be ruled out under LiM, at the cost of ruling out several other types. It is unlikely that this is a plausible scenario in empirical applications. As will be outlined below, justifying PM over LiM becomes even more difficult as the number of instruments increases.

Consider the three binary instrument setting with the three instruments $Z_1 \in \{0,1\}$, $Z_2 \in \{0,1\}$, and $Z_3 \in \{0,1\}$, and with support $\mathcal{Z} = \{(0,0,0),(0,0,1),(0,1,0),(1,0,0),(1,1,0),(0,1,1),(1,0,1),(1,1,1)\}$. Without imposing any restrictions, there are $2^8 = 256$ different response types, since there are eight different points of support of Z for which the potential treatment status is compared pair-wise. The eight different combinations of the instrument values result in $\binom{8}{2} = 28$ comparisons of potential treatments. LiM includes individuals who are a complier with respect to at least one of the instruments or a combination of instruments, but defiers for another instrument (or potentially multiple other instruments), as long as the treatment status when exposed to all instruments is at least as large as when exposed to none of the instruments. Imposing LiM ($P(D^{111} \geq D^{000}) = 1$ or $P(D^{111} \leq D^{000}) = 1$) rules out 64 of the initial 256 response types, allowing for a total of 192 possible types.

The maximum number of response types under PM is only 35 since it imposes more choice restrictions. PM imposes twelve restrictions in total that bring about $2^{12} = 4,096$ different versions of PM[4]. PM and LiM are nested in approximately 82% ($3,366/4,096 \approx 0.82$) of these cases. In all those instances, LiM is strictly weaker than PM. PM seems rather unrealistic when it is non-nested with LiM, which entails the remaining 18% of the versions of PM. These versions of PM only allow for either one, two or three additional response types excluded by LiM, at the cost of ruling out many other types that are included under LiM. In approximately 10% of all cases ($(730 - 324 - 12)/4,096$), one response type is allowed for under PM that is ruled out under LiM. In approximately 8% ($324/4,096$) of the cases, PM allows for two other response types. The maximum number of extra response types that PM allows for when non-nested with LiM is three, which occurs in 0.3% ($12/4,096$) of the possible combinations consistent with the PM assumption.

[4]An R-script for the response types that are allowed for under the different monotonicity assumptions in case of three binary instruments is available by the author upon request.

Table 2: Principal strata and the definition of the response types in case of two binary instruments and one binary treatment when LiM and PM are non-nested.

Type (T)	D^{11}	D^{10}	D^{01}	D^{00}	Notion	LiM (positive)	LiM (negative)	PM (Equation 2)	PM (Equation 3)	PM (Equation 4)	PM (Equation 5)
at	1	1	1	1	Always-taker	✓	✓	✓	✓	✓	✓
ec	1	1	1	0	Eager complier	✓		[✓]			
rc	1	0	0	0	Reluctant complier	✓					[✓]
$1c$	1	1	0	0	First instrument complier	✓			[✓]		
$2c$	1	0	1	0	Second instrument complier	✓				[✓]	
$1d$	1	0	1	1	First instrument defier	✓	✓			✓	✓
$2d$	1	1	0	1	Second instrument defier	✓	✓		✓		✓
ed	1	0	0	1	Eager defier	✓	✓				✓
rd	0	1	1	0	Reluctant defier	✓	✓	✓			
$d1$	0	1	0	0	Defier type 1	✓	✓	✓	✓		
$d2$	0	0	1	0	Defier type 2	✓	✓	✓		✓	
$d3$	0	1	1	1	Defier type 3		✓	(✓)			
$d4$	0	1	0	1	Defier type 4		✓		(✓)		
$d5$	0	0	1	1	Defier type 5		✓			(✓)	
$d6$	0	0	0	1	Defier type 6		✓				(✓)
nt	0	0	0	0	Never-taker	✓	✓	✓	✓	✓	✓

✓demonstrates the types allowed for under the respective forms of the monotonicity assumption.

(✓) denotes the one response type that is only allowed for under PM but excluded under positive LiM.

[✓] denotes the one response type that is only allowed for under PM but excluded under negative LiM.

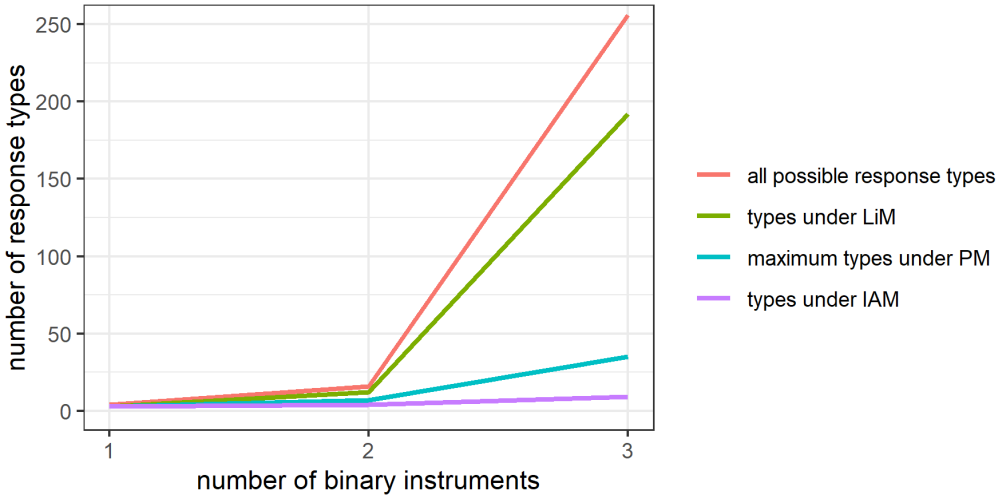


Figure 1: The maximum number of possible response types when one, two or three binary instruments are available under different versions of the monotonicity assumption. This figure shows that when more than one binary instrument is available, LiM imposes far fewer choice restrictions on the response types present in the population.

The number of defier types increases rapidly with the number of available binary instruments, because of the existence of defiers with respect to either instrument. The total possible number of response types is given by 2^{2^k}. Under LiM, 75% of the response types are allowed for and 25% are ruled out, independent of the number of instruments, k. The combined compliers always consist of 25% of the total number of response types, meaning that $0.25 \cdot 2^{2^k}$ response types form the combined compliers. Calculating the number of response types under PM is more complicated since the number of response types depends on the signs of the choice restrictions. Every choice restriction that is imposed eliminates at most 25% of the response types. The number of choice restrictions imposed by PM when k instruments are available equals $k \cdot 2^{k-1} = \sum_{i=1}^{k} \binom{k}{i-1} \cdot (k-i-1)$.

A graphic illustration of the restrictiveness of other forms of monotonicity compared to LiM is given in Figure 1. This figure depicts the maximum number of types allowed under each monotonicity assumption. It clearly demonstrates the advantage of imposing the LiM assumption as the number of available instruments increases with respect to the response types that are allowed for. PM forces the researcher to make a choice between types. Another problem is that, depending on the types and the ordering of the propensity scores, some response types can lead to negative weights in the weighted average estimated by TSLS. This will be further outlined in Section 5.

3.2 Examples

A main advantage of LiM is that it allows for flexibility in the response types, while PM is more restrictive with respect to the co-existence of defier and complier types. LiM is more flexible as it allows for some individuals to switch into treatment while other individuals switch out of treatment, as long as all individuals are equally or more likely to take up treatment when exposed to all instruments compared to being exposed to none of the instruments (since $P(D^{1\ldots1\ldots1} \geq D^{0\ldots0\ldots0}) = 1$). There are several applications where LiM might be more plausible to assume than PM[5].

3.2.1 Returns to college

Mogstad et al. (2021) provide the example of returns to college to motivate the plausibility of PM over the IAM assumption. The treatment is whether or not an individual attends college. The two binary instruments are a subsidy that forms a financial incentive to increase educational attainment, and the distance to college. PM allows for the co-existence of both first instrument compliers and second instrument compliers (see Table 1). However, as pointed out by Goff (2020), PM can be formulated such that it allows for some individuals to attend college when it is cheap and far, but this rules out the existence of individuals who would attend college only when it is cheap and close. Hence it does not allow for the co-existence of defier type 2 and reluctant compliers (see Table 3 for these two response types), because PM requires either $P(D^{11} \geq D^{01}) = 1$ or $P(D^{11} \leq D^{01}) = 1$. Unlike PM, LiM allows for these response types to co-exist, as it only imposes $P(D^{11} \geq D^{00}) = 1$.

Now consider the returns to college example with the distance to college as a first instrument and receiving personalized information about the likely returns to attending college as a second instrument. The personalized information might be provided by some kind of aptitude test, or by private guidance mentoring. One might consider imposing PM with the choice restrictions as in (1) and the corresponding response types from Table 1 if one believes that both instruments make college attendance always weakly more likely and that there are no defiers with respect to these instruments. Yet one can think of situations where these choice restrictions do not hold. For some students, an aptitude test might encourage them to attend college, while other students might be discouraged by the personalized information they obtain. The former group of individuals are compliers with respect to the information instrument, while the latter group are

[5]When instruments are of different treatment arms (consider for example preference-based instruments), then PM always is as plausible as LiM since there is no comparison between D^{1000} and D^{1100} since D^{1100} cannot occur. It is also not possible to estimate the CC-LATE in this setting, since there are no observations for which all instruments equal one. As mentioned by Mogstad et al. (2021), since the instruments are opposite treatment arms, their covariance is smaller than zero, and negative weights are possible.

Table 3

Type	D^{11}	D^{10}	D^{01}	D^{00}	Notion
rc	1	0	0	0	Reluctant complier
$d2$	0	0	1	0	Defier type 2

Under LiM, second instrument compliers and defiers of type 2 can co-exist since they satisfy the restriction $P(D^{11} \geq D^{00}) = 1$. Under PM, second instrument defiers and defiers of type 2 cannot co-exist because it requires either $P(D^{11} \geq D^{01}) = 1$ or $P(D^{11} \leq D^{01}) = 1$.

Table 4

Type	D^{11}	D^{10}	D^{01}	D^{00}	Notion
$2c$	1	0	1	0	Second instrument complier
$2d$	1	1	0	1	Second instrument defier

Under LiM, second instrument compliers and second instrument defiers can co-exist since they satisfy the restriction $P(D^{11} \geq D^{00}) = 1$. Under PM, second instrument compliers and second instrument defiers cannot co-exist because it requires either $P(D^{00} \geq D^{01}) = 1$ or $P(D^{00} \leq D^{01}) = 1$.

defiers with respect to this instrument. PM rules out the co-existence of these second instrument complier and second instrument defier types. LiM, however, allows for the co-existence of these response types.

More specifically, consider a setting with individuals who would defy going to college based on the incentive created by the personalized information, but who would go to college when they both receive the personalized information and live close to college. These individuals are defiers with respect to the information instrument, but respond more strongly to the distance instrument. Under LiM, these individuals can co-exist with the second instrument compliers who always attend college when they are encouraged by the aptitude test (see Table 4 for these two response types).

3.2.2 Twinning and same-sex instrument

Two instruments that are commonly used for exogenous variation in household size are the twinning instrument and the same-sex instrument. The twinning instrument equals one when the second and third child are twins, and the same-sex instrument equals one when the two firstborn in a family are of the same sex[6]. Angrist and Evans (1998) introduce the same-sex

[6]Mogstad et al. (2021) consider the same-sex instrument where one instrument is the gender of the first child and the other instrument the gender of the second child such that $P(D^{00} \geq D^{01}) = 1$, $P(D^{00} \geq D^{10}) = 1$, $P(D^{11} \geq D^{01}) = 1$, and $P(D^{11} \geq D^{10}) = 1$. This definition of the same-sex instrument would violate LiM, but

instrument based on the observation that parents generally have a preference for a mixed sibling-sex composition and compares this instrument to the twinning instrument. Since the sex of a child is basically randomly assigned, parents with two firstborn of the same sex are more likely to increase their household size with a third child. Moreover, since parents generally do not choose to have twins, twins at second and third birth can be seen as randomly assigning parents to have a household with three instead of two children. Thus, both instruments are commonly used to disentangle the effect of having a third child on female labor market outcomes.

While it is impossible to defy the twinning instrument, there might be defiers with respect to the same-sex instrument[7]. Generally, it is assumed that parents have a preference for siblings of opposite sex. Hence, if the two firstborn siblings are of the same sex, it is assumed that parents are more likely to have a third child. However, Dahl and Moretti (2008) find that the household size is larger when the firstborn is a girl and that boys are favored over girls in the United States. Moreover, De Chaisemartin (2017) mentions that the 2012 Peruvian wave of the Demographic and Health Surveys shows that 1.8% of the women with mixed firstborn composition and three children or more would have preferred either two boys or two girls retrospectively.

In these settings, LiM might be more plausible to assume than PM. Consider the two binary instruments where Z_1 is equal to one when the second and third children are twins, and Z_2 equals one when the two firstborn are of the same sex. Some parents might not have a third child when the two firstborn are boys, while they might have a third child when the two firstborn are mixed sex. These parents defy the same-sex instrument and can be considered second instrument defiers. Under PM, second instrument defiers cannot co-exist with second instrument compliers. Unlike PM, LiM allows for this situation[8]. That is, LiM allows for parents who do not have a third child when the two firstborn are of the same sex, but who would have had a third child otherwise.

---

can be re-written to a single instrument such that LiM holds: Define the instrument such that it equals one when the first two children are of the same sex, and zero when they are of opposite sexes.

[7]Note that some types are ruled out since defying the twinning instrument is impossible. Let the twinning instrument be Z_2. Then there are no never-takers or defiers of type 1.

[8]It should be noted that the CC-LATE uses observations for which either $Z_1 = 0$ and $Z_2 = 0$ or $Z_1 = 1$ and $Z_2 = 1$. In this example, the latter corresponds to observations of families where the two firstborn are of the same sex and the second and third children are twins. This means that for twins of the opposite sex, it matters which twin is born first.

4 Extension of the CC-LATE to more than two instruments

Suppose we have $k > 2$ binary instruments that all satisfy the LATE assumptions[9]. It can be shown that

$$E(Y^1 - Y^0|T \in cc) = \frac{E\left(Y|Z_1 = 1, \ldots, Z_k = 1\right) - E\left(Y|Z_1 = 0, \ldots, Z_k = 0\right)}{E\left(D|Z_1 = 1, \ldots, Z_k = 1\right) - E\left(D|Z_1 = 0, \ldots, Z_k = 0\right)},$$

where cc is the set of individuals who comply with at least one of the instruments or a combination thereof, while not defying any of the instruments when the other instrument values are all equal to zero or all equal to one. A great advantage of this parameter is that it is robust to the presence of many different defier types. More specifically, it allows for all defier types for which $P(D^{11...1} = D^{00...0}) = 1$.

Proof in Appendix D.

The size of the group of combined compliers is given by

$$E\left(D|Z_1 = 1, \ldots, Z_k = 1\right) - E\left(D|Z_1 = 0, \ldots, Z_k = 0\right)^{10\ 11}. \tag{6}$$

Increasing the number of instruments used to estimate the CC-LATE generally increases the combined complier population. However, increasing the number of instruments might decrease estimation precision since estimation is based only on data where $Z = (0.., 0..., 0)$ and $Z = (1.., 1..., 1)$. Estimating the CC-LATE is challenging if $P(D|Z_1 = 1, \ldots, Z_k = 1)$ is very small. This potentially gives a trade-off between the estimation precision and the size of the complier population.

Whether the expression given by (6) is small depends on the response type proportions, $\pi_t = P(T = t)$. The response type that an individual belongs to cannot be observed from the data. Instead, an individual can belong to several response groups based on the observed values of D and Z (see Table 5). Table 6 depicts the observed subgroups and the respective response types an individual can belong to under LiM in case of two binary instruments. The observed conditional treatment probabilities can then be assigned to the response type proportions as depicted in Table 7. Denote the combined compliers as the set $cc \equiv \{ec, rc, 1c, 2c\}$. Then

[9]Suppose we observe a multi-valued discrete instrument M that satisfy the LATE assumptions. For easy exposition assume that M takes values $0, 1, 2$. Define $Z_1 = I(M > 0)$ and $Z_2 = I(M > 1)$. As it is not possible to have $Z_1 = 0$ and $Z_2 = 1$, the Z_2 compliers and the eager compliers are ruled out in this case. Moreover, IAM automatically holds.

[10]It is worth noting that if $E\left(D|Z_1 = 1, \ldots, Z_k = 1\right) - E\left(D|Z_1 = 0, \ldots, Z_k = 0\right) > 0$ indicates positive monotonicity while $E\left(D|Z_1 = 1, \ldots, Z_k = 1\right) - E\left(D|Z_1 = 0, \ldots, Z_k = 0\right) < 0$ indicates negative monotonicity.

[11]The expression follows from

$$P(D_i^{11...1} > D_i^{00...0}) = P(D_i^{11...1} - D_i^{00...0} = 1) = E(D_i^{11...1} - D_i^{00...0}) = E(D_i^{11...1}) - E(D_i^{00...0}).$$

Table 5: Observed subgroups and response types with two binary instruments.

Observed subgroups	Response types
$\{i : Z_{1i} = 1, Z_{2i} = 1, D_i = 1\}$	Subject i belongs to $at,\ ec,\ rc,\ 1c,\ 2c,\ 1d,\ 2d,$ or ed
$\{i : Z_{1i} = 1, Z_{2i} = 1, D_i = 0\}$	Subject i belongs to $rd,\ d1,\ d2,\ d3,\ d4,\ d5,\ d6,$ or nt
$\{i : Z_{1i} = 1, Z_{2i} = 0, D_i = 1\}$	Subject i belongs to $at,\ ec,\ 1c,\ 2d,\ rd,\ d1,\ d3,$ or $d4$
$\{i : Z_{1i} = 1, Z_{2i} = 0, D_i = 0\}$	Subject i belongs to $rc,\ 2c,\ 1d,\ ed,\ d2,\ d5,\ d6,$ or nt
$\{i : Z_{1i} = 0, Z_{2i} = 1, D_i = 1\}$	Subject i belongs to $at,\ ec,\ 2c,\ 1d,\ rd,\ d2,\ d3,$ or $d5$
$\{i : Z_{1i} = 0, Z_{2i} = 1, D_i = 0\}$	Subject i belongs to $rc,\ 1c,\ 2d,\ ed,\ d1,\ d4,\ d6,$ or nt
$\{i : Z_{1i} = 0, Z_{2i} = 0, D_i = 1\}$	Subject i belongs to $at,\ 1d,\ 2d,\ ed,\ d3,\ d4,\ d5,$ or $d6$
$\{i : Z_{1i} = 0, Z_{2i} = 0, D_i = 0\}$	Subject i belongs to $ec,\ rc,\ 1c,\ 2c,\ rd,\ d1,\ d2,$ or nt

Table 6: Observed subgroups and response types under LiM with two binary instruments.

Observed subgroups	Response types
$\{i : Z_{1i} = 1, Z_{2i} = 1, D_i = 1\}$	Subject i belongs to $at,\ ec,\ rc,\ 1c,\ 2c,\ 1d,\ 2d,$ or ed
$\{i : Z_{1i} = 1, Z_{2i} = 1, D_i = 0\}$	Subject i belongs to $rd,\ d1,\ d2,$ or nt
$\{i : Z_{1i} = 1, Z_{2i} = 0, D_i = 1\}$	Subject i belongs to $at,\ ec,\ 1c,\ 2d,\ rd,$ or $d1$
$\{i : Z_{1i} = 1, Z_{2i} = 0, D_i = 0\}$	Subject i belongs to $rc,\ 2c,\ 1d,\ ed,\ d2,$ or nt
$\{i : Z_{1i} = 0, Z_{2i} = 1, D_i = 1\}$	Subject i belongs to $at,\ ec,\ 2c,\ 1d,\ rd,$ or $d2$
$\{i : Z_{1i} = 0, Z_{2i} = 1, D_i = 0\}$	Subject i belongs to $rc,\ 1c,\ 2d,\ ed,\ d1,$ or nt
$\{i : Z_{1i} = 0, Z_{2i} = 0, D_i = 1\}$	Subject i belongs to $at,\ 1d,\ 2d,$ or ed
$\{i : Z_{1i} = 0, Z_{2i} = 0, D_i = 0\}$	Subject i belongs to $ec,\ rc,\ 1c,\ 2c,\ rd,\ d1,\ d2,$ or nt

Table 7: Observed conditional treatment probabilities and corresponding principal strata proportions with $\pi_t = P(T = t)$ where $t = at, rc, ec, 1c, 2c, 1d, 2d, ed, rd, d1, d2, nt$ under LiM with two binary instruments.

Observed conditional treatment probabilities	Response group proportions
$P(D = 1\|Z_1 = 1, Z_2 = 1)$	$\pi_{at} + \pi_{ec} + \pi_{rc} + \pi_{1c} + \pi_{2c} + \pi_{1d} + \pi_{2d} + \pi_{ed}$
$P(D = 0\|Z_1 = 1, Z_2 = 1)$	$\pi_{rd} + \pi_{d1} + \pi_{d2} + \pi_{nt}$
$P(D = 1\|Z_1 = 1, Z_2 = 0)$	$\pi_{at} + \pi_{ec} + \pi_{1c} + \pi_{2d} + \pi_{rd} + \pi_{d1}$
$P(D = 0\|Z_1 = 1, Z_2 = 0)$	$\pi_{rc} + \pi_{2c} + \pi_{1d} + \pi_{ed} + \pi_{d2} + \pi_{nt}$
$P(D = 1\|Z_1 = 0, Z_2 = 1)$	$\pi_{at} + \pi_{ec} + \pi_{2c} + \pi_{1d} + \pi_{rd} + \pi_{d2}$
$P(D = 0\|Z_1 = 0, Z_2 = 1)$	$\pi_{rc} + \pi_{1c} + \pi_{2d} + \pi_{ed} + \pi_{d1} + \pi_{nt}$
$P(D = 1\|Z_1 = 0, Z_2 = 0)$	$\pi_{at} + \pi_{1d} + \pi_{2d} + \pi_{ed}$
$P(D = 0\|Z_1 = 0, Z_2 = 0)$	$\pi_{ec} + \pi_{rc} + \pi_{1c} + \pi_{2c} + \pi_{rd} + \pi_{d1} + \pi_{d2} + \pi_{nt}$

$E(D|Z_1 = 1, \ldots, Z_k = 1) = \pi_{at} + \pi_{cc} + \pi_{1d} + \pi_{2d} + \pi_{ed}$ is small if either the proportion of always-takers is small or the proportion of combined compliers is small. The latter can occur when only few individuals were assigned to all instruments. It might also occur if the instruments are weak, meaning that not many individuals respond to the instruments combined. $E(D|Z_1 = 1, \ldots, Z_k = 1)$ being large enough can therefore be seen as a relevance assumption of all instruments being strong.

Concluding this section, we can say that the CC-LATE is most likely to be useful in applications where the instruments are all strong, the number of instruments is relatively small, the sample size is very large, or observations with both $Z = (0.., 0..., 0)$ and $Z = (1.., 1..., 1)$ occur frequently. When there are only few observations for both $Z = (0.., 0..., 0)$ and $Z = (1.., 1..., 1)$, one might consider increasing the range of points included at the outer support of Z to decrease the variance at the cost of introducing some bias[12]. One might also consider discarding instruments that generate too few compliers.

5 Comparison of the CC-LATE to other estimands

The CC-LATE estimand is given by

$$\beta = \frac{E\left(Y|Z_1 = 1, \ldots Z_k = 1\right) - E\left(Y|Z_1 = 0, \ldots Z_k = 0\right)}{E\left(D|Z_1 = 1, \ldots Z_k = 1\right) - E\left(D|Z_1 = 0, \ldots Z_k = 0\right)}$$

when multiple binary instruments are available, and it is given by

$$\beta = \frac{E\left(Y \mid Z_1 = 1, Z_2 = 1\right) - E\left(Y \mid Z_1 = 0, Z_2 = 0\right)}{E\left(D \mid Z_1 = 1, Z_2 = 1\right) - E\left(D \mid Z_1 = 0, Z_2 = 0\right)}$$

in the case of two binary instruments.

When two binary instruments, Z_1 and Z_2, satisfy the standard assumptions including the IAM assumption, the Imbens and Angrist (1994) LATE estimands using each instrument separately are

$$\beta_1 = \frac{E\left(Y \mid Z_1 = 1\right) - E\left(Y \mid Z_1 = 0\right)}{E\left(D \mid Z_1 = 1\right) - E\left(D \mid Z_1 = 0\right)} \text{ and } \beta_2 = \frac{E\left(Y \mid Z_2 = 1\right) - E\left(Y \mid Z_2 = 0\right)}{E\left(D \mid Z_2 = 1\right) - E\left(D \mid Z_2 = 0\right)},$$

and the corresponding estimators $\widehat{\beta}_1$ and $\widehat{\beta}_2$ just replace the above expectations with sample averages. Let $\widehat{\beta}_1$ and $\widehat{\beta}_2$ be the estimated LATEs using Z_1 and Z_2 as instruments, respectively. Under standard assumptions, $\widehat{\beta}_1$ consistently estimates β_1, the average treatment effect among all first instrument compliers, and similarly $\widehat{\beta}_2$ consistently estimates β_2, the average treatment effect among all second instrument compliers. The denominators of these expression equal the probability of first instrument and second instrument compliers, respectively. The denominator

[12] This is comparable to the bias-variance trade-offs in Regression Discontinuity Designs.

of the CC-LATE estimand is always greater than or equal to the denominators of either β_1 or β_2 since it additionally includes eager compliers and reluctant compliers.

Imbens and Angrist (1994) show that when combining multiple instruments with TSLS under the IAM assumption, imposing choice homogeneity and using $g(Z)$ as an instrument, then TSLS gives a weighted average of the pairwise LATEs:

$$\alpha_g^{IV} = \sum_{k=1}^{K} \lambda_k \cdot E[Y_i(1) - Y_i(0) | D_i(z_k) = 1, D_i(z_{k-1}) = 0],$$

with weights

$$\lambda_k = \frac{(P(z_k) - P(z_{k-1})) \cdot \sum_{l=k}^{K} \pi_l \cdot (g(z_l) - E[g(Z)])}{\sum_{m=1}^{K} (P(z_m) - P(z_{m-1})) \cdot \sum_{l=m}^{K} \pi_l \cdot (g(z_l) - E[g(Z)])},$$

where $\pi_k = Pr(Z = z_k)$, $P(z_k) = E[D_i | Z_i = z_k]$, and the support of Z ordered such that if $l < m$, then $P(z_l) \leq P(z_m)$. The weights sum up to one. To guarantee positive weights, Imbens and Angrist (1994) additionally assume that $J(Z)$, the scalar instrument constructed from Z, depends on the propensity score $P(Z)$ in a monotone way[13]

Mogstad et al. (2021) show that under PM, TSLS gives a weighted average of the LATEs for the response types, g, in the population other than the always-takers and never-takers:

$$\beta_{\text{TSLS}} = \sum_{g \in \mathcal{G}: \mathcal{C}_g \neq \emptyset} \omega_g \cdot E[Y_i(1) - Y_i(0) | G_i = g], \tag{7}$$

with weights

$$\omega_g = P(G_i = g) \sum_{k=2}^{K} (\mathbb{1}[k \in \mathcal{C}_g] - \mathbb{1}[k \in \mathcal{D}_g]) \left(\frac{\text{Cov}(D_i, \mathbb{1}[p(Z_i) \geq p(z^k)])}{\text{Var}(p(Z_i))} \right)$$

with $\mathcal{C}_g$ and $\mathcal{D}_g$ being the sets of integers k for which a certain group type responds to the change from z^{k-1} to z^k as a complier or defier, with $\{z_1, ..., z^k\}$ the points of the instrument support ordered by the propensity scores, $p(z^1), ..., p(z^k)$. The weights sum to one. A drawback of this estimand is that its interpretation is not straightforward for two reasons: The weights are counter-intuitive and the LATEs of defier types might enter the weighted average. As is evident from the expression, negative weights can occur either if $\text{Cov}(D_i, \mathbb{1}[p(Z_i) < p(z^k)]) \leq 0$ or if $\mathbb{1}[k \in \mathcal{C}_g] - \mathbb{1}[k \in \mathcal{D}_g] = -1$. The latter expression can lead to negative weights if $\mathcal{D}_g \neq \emptyset$. When PM allows for both first instrument compliers and second instrument compliers, $\mathcal{D}_g \neq \emptyset$ always occurs for either one of these two types. So a negative weight on the LATE for one of these complier groups is generally a cause for concern when performing TSLS under PM. Even if the resulting weight is non-negative, the magnitude of the weight will be distorted if $\mathcal{D}_g \neq \emptyset$. Interpreting the TSLS estimand gets even more challenging when more than two instruments

[13] Heckman et al. (2006) show that the weights are always positive when $P(Z)$ is the instrument. Thus, the weights are always positive when the first stage of TSLS is fully saturated, since in this case $J(Z) = P(Z)$.

are available. The instruments generate a variety of different complier and defier types in this case. Consequently, there are many possibilities of having two-way flows for some change in the instrument values. Next consider the LATEs in the weighted average. The interpretation of the TSLS estimand depends on the LATEs of the response types present in the population, which is not straightforward in case of multiple instruments. A cause for concern is that $\mathcal{D}_g \neq \emptyset$ generally holds for defier types, causing these types to enter the weighted average in Equation (7).

An attractive property of the CC-LATE is that it always gives the effect in the population of combined compliers. The CC-LATE is robust to the many defier types that might exist under LiM. Moreover, it is not concerned with negative weights. The CC-LATE estimand can be interpreted as

$$\beta_{\text{CC-LATE}} = \sum_{g \in cc} \omega_g \cdot E[Y_i(1) - Y_i(0) | G_i = g],$$

with weights corresponding to the relative sizes of the complier groups:

$$\omega_g = P(G_i = g).$$

If PM and LiM are non-nested as discussed in Section 3, then it might not be possible to obtain an unbiased estimate of the CC-LATE if PM is true. Nevertheless, the CC-LATE parameter can still be more interesting to estimate than the TSLS parameter, because it might be close to the true LATE for the combined complier population (see Section 8 for a more detailed examination of the estimand under violation of LiM). Especially since the number of response types that are allowed for under PM but violating LiM are very few, as discussed previously in Section 3. However, when PM is violated, one should be careful with interpreting the TSLS parameter due to defier types entering the equation. This means that the weight can be negative, even if $\text{Cov}(D_i, \mathbb{1}[p(Z_i) < p(z^k)]) > 0$, which might even lead to the TSLS estimate having an opposite sign than the true ATE.

Frölich (2007) considers multiple instrumental variables with covariates included nonparametrically. If D_i follows an index structure and under standard assumptions including the IAM assumption which heavily restricts choice heterogeneity, he derives the following LATE:

$$E[Y^1 - Y^0 | \tau = c] = \frac{\int (E[Y|X=x, p(Z,X) = \bar{p}_x] - E[Y|X=x, p(Z,X) = \underline{p}_x]) f_x(x) dx}{\int (E[D|X=x, p(Z,X) = \bar{p}_x] - E[D|X=x, p(Z,X) = \underline{p}_x]) f_x(x) dx}$$

where $\bar{p}_x = max_z p(z,x)$ and $\underline{p}_x = min_z p(z,x)$. Similar to the CC-LATE, the estimation is based on the two subgroups of observations where $Z = (0.., 0..., 0)$ and $Z = (1.., 1..., 1)$. The interpretation of this estimand differs in that it considers the largest complier group, whereas the CC-LATE considers all individuals that respond to any instrument or combination thereof. From the results of Imbens and Angrist (1994), one can show that $\frac{E[Y|Z=z_K] - E[Y|Z=z_0]}{E[D|Z=z_K] - E[D|Z=z_0]} = E[Y(1) - Y(0) | D(z_K) \neq D(z_0)]$ (see Appendix E) which equally can be interpreted as the effect in the

largest group of compliers, having the same interpretation as the estimand for multiple binary instruments as proposed by Frölich (2007).

Goff (2020) derives the "all compliers LATE" (ACL) under a special form of PM which he refers to as vector monotonicity (VM). He shows that the ACL can be re-written to a weighted average over the treatment effects of the specific combined complier groups, $g \in \mathcal{G}$:

$$E[Y_i(1) - Y_i(0)|C_i = 1] = \sum_{g \in \mathcal{G}} \frac{P(G_i = g)E[c(g, Z_i)]}{E[c(G_i, Z_i)]} \cdot E[Y_i(1) - Y_i(0)|G_i = g] \tag{8}$$

where $C_i = c(G_i, Z_i) = 1$ if a unit i belongs to a group of the all compliers. Identification of the ACL is then possible for specific choices of the function $c(g, z)$. Only in rare cases does the TSLS estimator recover the ACL and Goff (2020) proposes a different estimator that is similar in construction to the TSLS estimator. He further shows that Equation (8) can be re-written to a single Wald estimand:

$$E[Y_i(1) - Y_i(0)|C_i = 1] = \frac{E[Y_i|Z_i = \bar{Z}] - E[Y_i|Z_i = \underline{Z}]}{E[D_i|Z_i = \bar{Z}] - E[D_i|Z_i = \underline{Z}]}$$

where $\bar{Z} = (1, 1, ..., 1)'$ and $\underline{Z} = (0, 0, ..., 0)'$. Obviously, the denominator should be nonzero and it should hold that $P(Z_i = \bar{Z}) > 0$ and $P(Z_i = \underline{Z}) > 0$.

As the name suggests, the "all compliers" LATE concerns individuals who are compliers in the sense that they respond to the instruments in some way. Under VM, the ACL gives the effect for those individuals who are neither always-takers nor never-takers. In the setting with two binary instruments, the interpretation of the CC-LATE is coincides with the interpretation of the ACL in that it estimates the effect for those individuals who are a complier with respect to one of the instruments without defying any of the other instruments. In this case, the combined complier population coincides with the all complier population. However, the CC-LATE is derived under a much weaker monotonicity assumption that allows for more choice heterogeneity and a rich co-existence of defier types. In the setting with three or more binary instruments, the combined complier population considered by the CC-LATE contains complier types that are ruled out under the VM assumption. Consequently, the CC-LATE gives the LATE for a larger complier population. The ACL is not necessarily identified in cases where VM is violated, but LiM still holds.

6 Empirical application - Impact of learning HIV status

In this section we illustrate the results of our preceding sections by estimating the effect of learning one's HIV status on protective behavior to prevent spread of the disease. Learning about a negative HIV test result might motivate individuals to further protect themselves, while learning about a positive result might motivate individuals to not spread the disease. The effect

Table 8

Type	D^{11}	D^{10}	D^{01}	D^{00}	Notion
$2c$	1	0	1	0	Distance instrument complier
$2d$	1	1	0	1	Distance instrument defier

Under LiM, distance instrument compliers and distance instrument defiers can co-exist since they satisfy the restriction $P(D^{11} \geq D^{00}) = 1$. Under PM, distance instrument compliers and distance instrument defiers cannot co-exist because it requires either $P(D^{00} \geq D^{01}) = 1$ or $P(D^{00} \leq D^{01}) = 1$.

of learning test results on the spread of the disease is very important from a policy perspective. Since learning the test results is an individual choice, selection bias is a serious threat in this application. Thornton (2008) investigates the effect of knowledge about the own HIV status on the purchase of contraceptives in rural Malawi, dealing with selection issues by instrumenting the endogenous decision of learning HIV test results with a financial incentive offered in the form of cash to pick up the test result and with the distance to the recommended HIV center.

6.1 Data

For all our analyses we use the sample as in Thornton (2008). The complete-case sample contains HIV-positive and negative individuals in Balaka and Rumphi who had sex and got tested for HIV in 2004 and took part in a follow-up survey in 2005. Similar to Thornton (2008), we consider four different outcomes. The outcomes are (1) whether or not an individual bought condoms at the follow-up survey that took place two months after testing, (2) how many condoms the individual bought at the follow-up survey, (3) if the individual reported buying condoms between getting tested and the follow-up survey, and (4) whether the individual reported having sex between getting tested and the follow-up survey. The treatment is whether or not the individual obtained the HIV test results and hence is aware of their HIV status.

We consider three instruments. The first instrument equals one when an individual received *any cash* incentive and zero otherwise. The second instrument is a *distance* incentive that equals one when distance to HIV test center is smaller than 1.5km and zero otherwise[14]. We further construct a third instrument, *above median cash* incentive, that equals one if the individual received an amount of the cash incentive above the median amount, and zero otherwise.

6.2 Motivation for LiM and the CC-LATE

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 HIV application with the *any cash* and *distance* instrument, $\bar{D}^{00} = 0.388$, $\bar{D}^{10} = 0.805$, $\bar{D}^{01} = 0.392$, and $\bar{D}^{11} = 0.832$[15]. This implies that the ordering of PM as in Equation (1) in Section 3 is consistent with the data, leading to the response types as listed in Table 1 in Section 2. It is worth noting that in this case, TSLS estimates a weighted average of the LATEs of the combined complier groups. TSLS estimates the weighted average over the combined compliers, that is, those individuals who comply either to the distance or to the cash incentive or to both or to either one of them. The CC-LATE estimates a single LATE for the combined complier population.

We argue that assuming LiM is more plausible in this application than assuming PM and estimating the CC-LATE gives a more policy relevant estimate as the one obtained by combining the instruments with TSLS. First of all, LiM is more plausible regarding the response types potentially present in the population. Living close to the recommended HIV center might encourage some individuals to learn their HIV status due to the small effort of traveling to the center. Meanwhile, it might discourage other individuals who would feel too embarrassed to visit an HIV center in their neighborhood out of fear of being recognized. These individuals are defiers with respect to the instrument for the proximity of an HIV center and defy learning their HIV status when living close to the recommended HIV center. However, they could be willing to learn their status when additionally receiving a financial incentive. In this case, the financial incentive outweighs the cost of being recognized and feeling embarrassed, pushing them into the treatment of learning their HIV status. PM would be violated if, next to these individuals, there exist individuals who always comply with the proximity instrument. LiM, however, would still hold since it allows for the co-existence of proximity instrument compliers and proximity instrument defiers (see Table 8 for these two response types). LiM only requires that when individuals receive cash and live close to a center, they do not defy learning their HIV test results. As pointed out by Thornton (2008), social stigma can prevent individuals from learning their HIV status. She finds that social barriers can be lifted by financial incentives, as the cash provides an excuse for visiting the HIV test center.

Secondly, even if PM is satisfied, we cannot reject negative weights which makes interpretation of TSLS problematic. We use Mogstad et al. (2021)'s approach to check whether the weights remain positive under PM when IAM is violated through the presence of both Z_1 and

[14] Technically, the location of the test centers was randomized and not the distance to the test centers.

[15] In case of three instruments, $\bar{D}^{000} = 0.39$, $\bar{D}^{100} = 0.70$, $\bar{D}^{010} = 0.39$, $\bar{D}^{110} = 0.75$, $\bar{D}^{101} = 0.91$, and $\bar{D}^{111} = 0.92$. This implies the following choice restrictions: $P(D^{100} \geq D^{000}) = 1$, $P(D^{010} \geq D^{000}) = 1$, $P(D^{110} \geq D^{100}) = 1$, $P(D^{101} \geq D^{100}) = 1$, $P(D^{110} \geq D^{010}) = 1$, $P(D^{111} \geq D^{110}) = 1$, and $P(D^{111} \geq D^{101}) = 1$. The data provides no information on $P(D^{000} \geq D^{001})$, $P(D^{001} \geq D^{011})$, $P(D^{010} \geq D^{011})$, $P(D^{001} \geq D^{101})$, and $P(D^{011} \geq D^{111})$ since there are no observations for $(z_1 = 0, z_2 = 0, z_3 = 1)$ and $(z_1 = 0, z_2 = 1, z_3 = 1)$.

Table 9: Testing for negative TSLS weights when both Z_1 and Z_2 compliers exist and IAM is relaxed to PM. Each column shows the coefficient from a regression of the column on the variable in the row including a constant. Standard errors in parentheses.

	(1)	(2)	(3)	(4)	(5)	(6)
	Got results	Got results	Got results	Any cash	Any cash	Distance
Any cash	0.425***					
	(0.032)					
Distance		0.024		0.003		
		(0.029)		(0.027)		
Median cash			0.303***		0.343***	-0.003
			(0.027)		(0.024)	(0.031)

Z_2 compliers. They are positive under a violation of this assumption if the correlation between the treatment and the instruments is positive and significant, and the partial correlation between the instruments is significant. We follow their approach and regress the treatment on each instrument separately. We also regress Z_1 on Z_2 and Z_3 separately, and Z_2 on Z_3. The results are in Table 9. The correlation between the *distance* instrument and the treatment is not significant (see Column (2) of Table 9). The partial correlation between the *above median cash* and *distance* instruments is also not positive (see Column (6) of Table 9). This indicates that TSLS might contain negative weights when the IAM assumption is replaced by the weaker PM assumption.

We perform two tests on the TSLS weights. We cannot reject the hypothesis that all weights are positive when performing TSLS with the two instruments, *any cash* instrument and *distance* instrument [16]. However, we also do not reject the hypothesis that one of the weights in the weighted average generated by TSLS is negative, finding a p-value of 0.207. One or more of the weights being negative would be extremely problematic for the interpretation of the TSLS estimates.

Finally, note that the CC-LATE parameter gives the average treatment effect for those individuals who change their treatment status when they are both offered a cash incentive and live close to the recommended HIV center. This makes the CC-LATE especially interesting, since the instruments considered here can be manipulated by the policy-maker. For instance, more HIV centers could be established and more cash incentives could be offered. Since the individual LATEs give the LATEs for the respective complier groups, examining the CC-LATE in combination with the single LATEs can be valuable from a policy perspective when deciding

[16] Using the *mivcausal* package (Lau and Torgovitsky, 2020), we obtain a p-value of 0.855 using 1000 repetitions in the bootstrap.

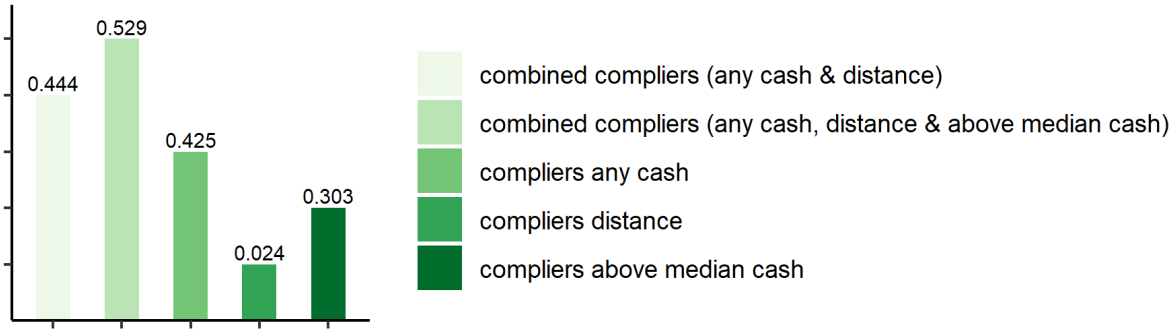


Figure 2: Shares of complier population for different instrument configurations.

whether the instruments should be offered jointly or separately.

6.3 Instrument distribution and complier share

The CC-LATE uses observations at the extremes of the support, that is, those observations for which all instrument values are zero and those for which all instrument values equal one. In the setting with the two instruments, *any cash* and *distance*, 43% of the observations are used to estimate the CC-LATE (see Table 10). In the setting with all three instruments, 27% of the observations are used to estimate the CC-LATE (see Table 10). Including the third instrument, *above median cash*, hence leads to a loss of 16% of total number of observations. Adding instruments always leads to the same or fewer amount of observations used for estimating the CC-LATE.

The probability of being a Z_1, Z_2 or Z_3 complier and the probability of being a combined complier in the two instrument setting are summarized in Figure 2. The share of compliers for the *distance* instrument is only 2.4%. Unsurprisingly, the largest complier share is reached with 52.9% when all three instruments are used. Adding an instrument to an existing set of instruments always makes the complier population at least as large. Since adding instruments never decreases the complier population while simultaneously decreasing the observations used for estimating the CC-LATE, there is a trade-off between the size of the complier population considered and the efficiency of the CC-LATE estimator.

6.4 Results

We estimate the effect of HIV learning on the four aforementioned outcomes with OLS, the CC-LATE estimator, and TSLS. Standard errors are robust and clustered at the village level. Controls are omitted[17]. The OLS seem to be downward biased (see Figure 3a). It might be that respondents who do not practice safe sex were more likely to choose to learn their HIV status,

[17]Since the instruments are randomized, omitting controls does not introduce bias.

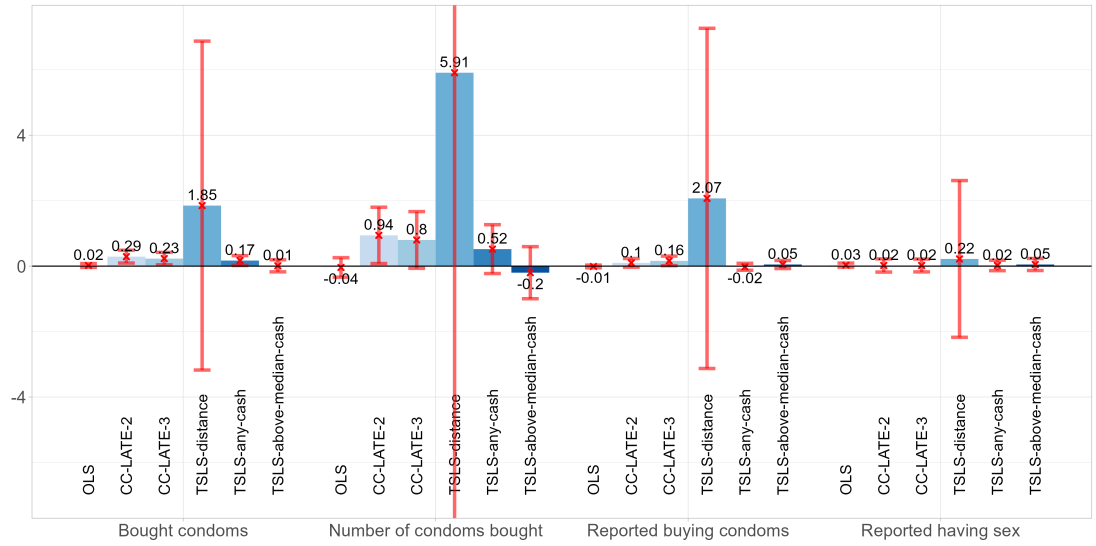
Table 10: Distribution of the instruments in the setting with two instruments and three instruments in the complete-case data.

	Z_1	Z_2	Z_3		
	any cash	*distance*	*above median cash*	nr. observations	% observations
two instruments	0	0		134	13%
	0	1		497	49%
	1	0		79	8%
	1	1		298	30%
total nr. of observations				1008	100%
observations used by CC-LATE				432	43%
three instruments	0	0	0	134	13%
	1	0	0	79	8%
	0	1	0	254	25%
	0	0	1	0	0%
	1	1	0	154	15%
	1	0	1	0	0%
	0	1	1	243	24%
	1	1	1	144	14%
total nr. of observations				1008	100%
observations used by CC-LATE				278	28%

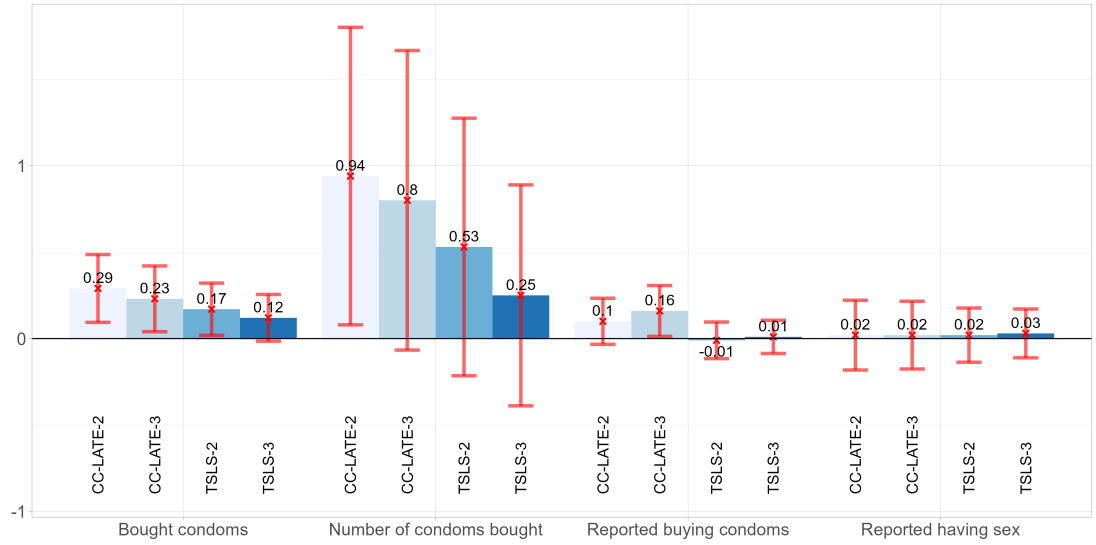
or it might be that individuals who do practice safe sex were less likely to choose to learn HIV status.

Fortunately, comparing the *CC-LATE-2* and *CC-LATE-3* estimates in Figure 3a, it seems that adding a third instrument does not influence the efficiency of the CC-LATE estimator much in this application (with the exception of the second outcome, *number of condoms bought*). The estimate decreases in magnitude when adding the *above median cash* instrument. This might be due to the fact that this instrument adds compliers that really need to be pushed towards compliance.

Figure 3a also gives the estimates obtained when using each instrument separately. Clearly, for all four outcome variables, the estimate obtained when using the *distance* instrument individually is larger in magnitude with much wider confidence intervals. The F-statistic of this instrument is rather small (approximately 3), making it a "weak" instrument. This is reflected in the estimates. By adding the *distance* instrument, observations at the extreme of the support are lost and therewith efficiency, while it does not add much to the complier share. Therefore, it also does not contribute much to external validity. Using only the *above median cash* instrument, a different sign for two out of the four outcomes is estimated. Not only can these differences be due to different complier populations, but using each instrument separately might result in less



(a) Comparison of CC-LATE estimates to the estimates resulting from using each instrument separately. The confidence intervals for *TSLS distance* for the outcomes "number of condoms bought" is $[-12.66, 24.48]$. Figure 4 in Appendix I excludes the estimate of the distance instrument from Figure (a) for easier comparison.



(b) Comparison of CC-LATE estimates to the estimates when combining the instruments with TSLS.

Figure 3: These figures show the CC-LATE and TSLS estimates for the four outcome variables. The treatment is whether an individual learned their HIV status. In the setting with two instruments (e.g., *CC-LATE-2*), *any cash* (if any financial incentive was received) and *distance* (HIV center within 1.5km distance was offered) are used as instruments. In the setting with three instruments (e.g., *CC-LATE-3*), *above median cash* (one if total incentive above median, zero otherwise) is added as an instrument. Cluster bootstrapped standard errors (1000 repetitions). 95% confidence intervals in red. The estimates can also be found in Tables 11 and 12 in Appendix H.

precise estimates. Combining the instruments might be more advantageous.

We now compare CC-LATE estimates with the estimates obtained when combining the instruments with TSLS as is typically done in literature. The estimates are depicted in Figure 3b[18]. Reassuringly, the confidence intervals of the CC-LATE estimates and TSLS estimates are comparable. The first outcome considered is whether an individual bought condoms at the follow-up survey. Individuals who got their test results were 23 percentage points more likely to buy condoms according to CC-LATE estimate with three instruments (*any cash*, *distance*, *above median cash*). This is 12 percentage points for TSLS with three instruments.

For the second outcome, the CC-LATE estimate indicates that individuals who learned their HIV status bought on average 0.8 condoms more at the follow-up survey. TSLS estimates that individuals who learned their status bought 0.28 condoms more on average. For the setting with two instruments, the CC-LATE estimate is not only larger in magnitude but also significant at the 10% level while the TSLS estimator is not significant at the 10% level. The estimates for the final outcome, "reported having sex", is not significant at the 10% level.

Interestingly, adding compliers who respond to the *above median cash* instrument leads to an increase in the CC-LATE estimate for the "reported buying condoms" outcome while it leads to a decrease for the "bought condoms" outcome. While the former outcome captures whether the respondent bought condoms between getting tested and the follow-up survey, the latter outcome captures whether the 30 cents they received at the end of the follow-up survey were subsequently used to buy the offered subsidized condoms. The difference in estimates for two and three instruments between these two outcomes may be explained by the fact that the individuals who had to be pushed to compliance by a stronger financial incentive might be lying when responding to the question whether they bought condoms before the follow-up survey. These individuals subsequently do not buy condoms since they would rather keep the money.

Overall, the CC-LATE estimates provide more evidence for protective behavior after learning one's HIV status compared to the TSLS estimates[19]. Differences in estimates can be attributed to differences in interpretation or a violation of the PM assumption. The weighted average estimated by TSLS might contain negative or distorted weights. Recall that we were not able to reject the hypothesis that one of the weights is negative. Moreover, if distance instrument defiers are present, then PM is violated and the weighted average contains the LATE of this defier type. The CC-LATE is robust to the presence of this defier type. While the response types considered

[18]See Figure Iin Appendix I for Figure 3a without the *distance* instrument to allow for easier comparison of the CC-LATE estimator with the LATEs of each instrument used separately.

[19]Note that the treatment is whether someone chose to learn their HIV status and does not account for differences in learning whether the individual tested positive or negative for HIV. We find similar effects in the subsample with individuals who test negative. The subsample with individuals who test positive is too small to draw meaningful conclusions.

are the same for PM and LiM in case of two instruments, differences in estimates in case of three instruments can be attributed to the fact that there are 64 combined complier types under LiM. If PM holds, the TSLS estimates five a weighted average of at most 35 response types.

7 Conclusion

TSLS is often used in empirical applications to combine multiple instruments. This paper has highlighted a number of problems with this approach as well as the restrictiveness of commonly invoked monotonicity assumptions. We introduce a more plausible monotonicity assumption which we refer to as LiM and we introduce the CC-LATE, an arguably more policy relevant causal parameter. The CC-LATE is defined for a larger complier population and is robust to the presence of a variety of defier types. Additionally, we apply our method to estimate the effect of learning one's HIV status on protective behavior. The CC-LATE estimates provided more evidence of protective behavior than the one obtained with TSLS. These findings have important policy implications. They indicate that programs encouraging learning one's HIV status using cash and distance incentives might slightly prevent the spread of the disease, even in the presence of distance instrument defiers.

8 Extensions

8.1 Violation of LiM

In this section, we consider identification when LiM is violated. Violation of this assumption not only introduces identification issues, but also reduces the power of the instruments which exacerbates the problem (Dahl et al., 2017). If LiM is violated, it can be shown for the setting with two binary instruments that

$$\beta = \frac{\pi_{cc}}{\pi_{cc} - \pi_{dd}} E(Y^1 - Y^0 | T \in cc) - \frac{\pi_{dd}}{\pi_{cc} - \pi_{dd}} E(Y^1 - Y^0 | T \in dd)$$

with cc the set of combined compliers, $cc \equiv \{ec, rc, 1c, 2c\}$, and dd the set of defiers that can never be pushed towards compliance and do not cancel out, $dd \equiv \{d3, d4, d5, d6\}$.

Proof in Appendix F.

This result can easily be extended to the setting with multiple binary instruments. In the setting with three or more instruments, the set dd contains those individuals who are a defier with respect to one of the instruments when the values of the other instruments are either all equal to zero or when they are all equal to one.

If the probability of being this type of defier is small, that is, π_{dd} is small, then more weight is given to $E(Y^1 - Y^0|T \in cc)$ such that the impact of these defiers will be small. The same holds when the average treatment effect for these defiers is negligible, that is, $E(Y^1 - Y^0|T \in dd)$ is very small compared to the effect in the combined compliers group, $E(Y^1 - Y^0|T \in cc)$. The presence of these defiers is problematic when they are many and/or their treatment effect is relatively large in magnitude. In this case, they will introduce a substantial bias. There are not many settings where it is likely that these types of defiers introduce a large amount of bias, especially since LiM already allows for the existence of a rich set of defiers. The CC-LATE is identified if $E(Y^1 - Y^0|T \in cc) = E(Y^1 - Y^0|T \in dd)$.

The CC-LATE under a violation of LiM is a weighted average of the ATE for the combined compliers and the ATE for the defier types that would have been ruled out under LiM with negative weight. This is comparable to the TSLS estimand which is a weighted average that potentially contains defier types and/or negative weights.

8.2 Bloom result

In a randomized trial with one-sided noncompliance there are no never-takers. For the setting with one binary instrument, Bloom (1984) shows that IV estimates the treatment effect on the treated in randomized trials with one-sided noncompliance,

$$\frac{E(Y_i|z_i = 1) - E(Y_i|z_i = 0)}{P(D_i = 1|z_i = 1)} = E(Y_{1i} - Y_{0i}|D_i = 1).$$

When there are two binary instruments, one-sided compliance means that

$$E(D_i|Z_1 = 0, Z_2 = 0) = P(D_i = 1|Z_{1i} = 0, Z_{2i} = 0) = \pi_{at} + \pi_{1d} + \pi_{2d} + \pi_{ed} = 0.$$

If compliance is only possible when both instruments are offered such that $Z_{1i} = 1, Z_{2i} = 1$, then the average treatment effect on the treated (ATT) is

$$E(Y_i^1 - Y_i^0|D_i = 1) = \frac{E(Y_i|Z_{1i} = 1, Z_{2i} = 1) - E(Y_i|Z_{1i} = 0, Z_{2i} = 0)}{P(D_i = 1|Z_{1i} = 1, Z_{2i} = 1)}.$$

Proof in Appendix G.

This result can easily be extended to the setting with more than two binary instruments if it holds that compliance is only possible when an individual is exposed to all instruments.

If one-sided compliance only holds for one of the instruments, Z_2, and compliance is only possible when both instruments are offered, then

$$E(Y_i^1 - Y_i^0|D_i = 1) = \frac{E(Y_i|Z_{1i} = 1, Z_{2i} = 1) - E(Y_i|Z_{2i} = 0)}{P(D_i = 1|Z_{1i} = 1, Z_{2i} = 1)}.$$

8.3 Characteristics of the complier groups

When multiple instrumental variables are available, each instrument identifies the LATE for those individuals who change their treatment status in response to a change in that specific instrument. As pointed out in Angrist and Pischke (2008), when treatment effects are heterogeneous, the LATEs might differ due to differences in complier populations. Characteristics of the different complier populations might explain some of the differences in the estimated effects. Furthermore, LATEs are criticized for their lack of external validity. Knowledge about the characteristics of the population for which the average treatment effect was estimated might be valuable when extending to other populations.

Suppose there is a binary variable, X, that equals one when an individual is male, and zero when an individual is female.

$$\begin{aligned}
&\frac{P(x_{1i}=1|D_i^{11...1} > D_i^{00...0})}{P(x_{1i}=1)} \\
&= \frac{P(D_i^{11...1} > D_i^{00...0}|x_{1i}=1)}{P(D_i^{11...1} > D_i^{00...0})} \\
&= \frac{E(D_i|Z_{1i}=1, Z_{2i}=1, ..., Z_{ki}=1, x_{1i}=1) - E(D_i|Z_{1i}=0, Z_{2i}=0, ..., Z_{ki}=0, x_{1i}=1)}{E(D_i|Z_{1i}=1, Z_{2i}=1, ..., Z_{ki}=1) - E(D_i|Z_{1i}=0, Z_{2i}=0, ..., Z_{ki}=0)}
\end{aligned}$$

Thus, we can obtain the relative likelihood of a combined complier being male through the first stage and the first stage for male individuals only.

References

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

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

Angrist, J. D. and Pischke, J.-S. (2008). Mostly Harmless Econometrics. In *Mostly Harmless Econometrics*. Princeton university press.

Bloom, H. S. (1984). Accounting for No-Shows in Experimental Evaluation Designs. *Evaluation Review*, 8(2):225–246.

Dahl, C. M., Huber, M., and Mellace, G. (2017). It's Never too LATE: A New Look at Local Average Treatment Effects with or without Defiers. *Discussion Papers on Business and Economics, University of Southern Denmark*, 2.

Dahl, G. B. and Moretti, E. (2008). The Demand for Sons. *The Review of Economic Studies*, 75(4):1085–1120.

De Chaisemartin, C. (2017). Tolerating Defiance? Local Average Treatment Effects without Monotonicity. *Quantitative Economics*, 8(2):367–396.

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

Goff, L. (2020). A Vector Monotonicity Assumption for Multiple Instruments. *arXiv preprint arXiv:2009.00553*.

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.

Huber, M., Lechner, M., and Wunsch, C. (2013). The Performance of Estimators based on the Propensity Score. *Journal of Econometrics*, 175(1):1–21.

Hull, P. (2018). Isolateing: Identifying Counterfactual-Specific Treatment Effects with Cross-Stratum Comparisons. *Available at SSRN 2705108*.

Huntington-Klein, N. (2020). Instruments with Heterogeneous Effects: Bias, Monotonicity, and Localness. *Journal of Causal Inference*, 8(1):182–208.

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

Kirkeboen, L. J., Leuven, E., and Mogstad, M. (2016). Field of Study, Earnings, and Self-Selection. *The Quarterly Journal of Economics*, 131(3):1057–1111.

Kolesár, M. (2013). Estimation in an Instrumental Variables Model with Treatment Effect Heterogeneity. *Unpublished*.

Lau, C. and Torgovitsky, A. (2020). mivcausal: A Stata Module for Testing the Hypothesis about the Signs of the 2SLS Weights.

Mogstad, M., Santos, A., and Torgovitsky, A. (2018). Using Instrumental Variables for Inference about Policy Relevant Treatment Parameters. *Econometrica*, 86(5):1589–1619.

Mogstad, M. and Torgovitsky, A. (2018). Identification and Extrapolation of Causal Effects with Instrumental Variables. *Annual Review of Economics*, 10:577–613.

Mogstad, M., Torgovitsky, A., and Walters, C. R. (2020). Policy Evaluation with Multiple Instrumental Variables. *Unpublished Working Paper*.

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.

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

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

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

Small, D. S., Tan, Z., Ramsahai, R. R., Lorch, S. A., and Brookhart, M. A. (2017). Instrumental Variable Estimation with a Stochastic Monotonicity Assumption. *Statistical Science*, 32(4):561–579.

Thornton, R. L. (2008). The Demand for, and Impact of, Learning HIV Status. *American Economic Review*, 98(5):1829–63.

Vytlacil, E. (2002). Independence, Monotonicity, and Latent Index Models: An Equivalence Result. *Econometrica*, 70(1):331–341.

Appendices

A Proof of Theorem 1

Assume our data consists of independent, identically distributed observations of the vector $(Y_i, D_i, Z_{1i}, Z_{2i})$ for individuals $i = 1, ..., n$. Define the following four variables:

$$R_{1i} = (1 - Z_{1i})(1 - Z_{2i}), \quad R_{2i} = Z_{1i}Z_{2i}, \quad R_{3i} = (1 - Z_{1i})Z_{2i}, \quad R_{4i} = Z_{1i}(1 - Z_{2i}).$$

Under SUTVA the observed treatment D_i assigned to an individual i can be written as

$$\begin{aligned} D_i &= (1 - Z_{1i})(1 - Z_{2i})D_i^{00} + Z_{1i}Z_{2i}D_i^{11} + (1 - Z_{1i})Z_{2i}D_i^{01} + Z_{1i}(1 - Z_{2i})D_i^{10} \\ &= D_i^{00}R_{1i} + D_i^{11}R_{2i} + D_i^{01}R_{3i} + D_i^{10}R_{4i}. \end{aligned}$$

Consider the denominator of the CC-LATE estimand:

$$\begin{aligned} E\left(D|Z_1 = 1, Z_2 = 1\right) - E\left(D|Z_1 = 0, Z_2 = 0\right) &= E(D|R_2 = 1) - E(D|R_1 = 1) \\ &= E(D_i^{11}|R_2 = 1) - E(D_i^{00}|R_1 = 1) \\ &= E(D_i^{11}) - E(D_i^{00}). \end{aligned}$$

Let $\pi_t = \Pr(T \in t), t = at, rc, ec, 1c, 2c, 1d, 2d, ed, rd, d1, d2, nt$ (see Table 1). We have

$$\begin{aligned} E(D_i^{00}) =& \sum_t E(D_i^{00}|T = t)\pi_t \\ =& \pi_{at} \cdot 1 + \pi_{rc} \cdot 0 + \pi_{ec} \cdot 0 + \pi_{1c} \cdot 0 + \pi_{2c} \cdot 0 + \pi_{1d} \cdot 1 + \pi_{2d} \cdot 1 + \pi_{ed} \cdot 1 + \pi_{rd} \cdot 0 + \pi_{d1} \cdot 0 \\ &+ \pi_{d2} \cdot 0 + \pi_{nt} \cdot 0 \\ =& \pi_{at} + \pi_{1d} + \pi_{2d} + \pi_{ed} \end{aligned}$$

and

$$\begin{aligned} E(D_i^{11}) =& \sum_t E(D_i^{11}|T = t)\pi_t \\ =& \pi_{at} \cdot 1 + \pi_{rc} \cdot 1 + \pi_{ec} \cdot 1 + \pi_{1c} \cdot 1 + \pi_{2c} \cdot 1 + \pi_{nt} \cdot 0 + \pi_{1d} \cdot 1 + \pi_{2d} \cdot 1 + \pi_{ed} \cdot 1 + \pi_{rd} \cdot 0 \\ &+ \pi_{d1} \cdot 0 + \pi_{d2} \cdot 0 \\ =& \pi_{at} + \underbrace{\pi_{rc} + \pi_{ec} + \pi_{1c} + \pi_{2c}}_{\pi_{cc}} + \pi_{1d} + \pi_{2d} + \pi_{ed} \\ =& \pi_{at} + \pi_{cc} + \pi_{1d} + \pi_{2d} + \pi_{ed}. \end{aligned}$$

It therefore follows that

$$E(D|Z_1 = 1, Z_2 = 1) - E(D|Z_1 = 0, Z_2 = 0) = E(D_i^{11}) - E(D_i^{00}) = \pi_{cc},$$

which is the probability of being any type of complier.

Let $\beta_i = Y_i^1 - Y_i^0$. Note that unlike the CC-LATE β, the term β_i is random. Under SUTVA the observed outcome Y can be written as

$$\begin{aligned} Y_i &= Y_i^1 D_i + Y_i^0(1 - D_i) = \beta_i D_i + Y_i^0 \\ &= \beta_i \left[D_i^{00} R_{1i} + D_i^{11} R_{2i} + D_i^{01} R_{3i} + D_i^{10} R_{4i} \right] + Y_i^0 \\ &= \beta_i D_i^{00} R_{1i} + \beta_i D_i^{11} R_{2i} + \beta_i D_i^{01} R_{3i} + \beta_i D_i^{10} R_{4i} + Y_i^0. \end{aligned}$$

Now, consider the numerator of the CC-LATE estimand,

$$\begin{aligned} E\left(Y|Z_1 = 1, Z_2 = 1\right) - E\left(Y|Z_1 = 0, Z_2 = 0\right) =& E(Y|R_2 = 1) - E(Y|R_1 = 1) \\ =& E(\beta_i D_i^{11} + Y_i^0|R_2 = 1) - E(\beta_i D_i^{00} + Y_i^0|R_1 = 1) \\ =& E(\beta_i D_i^{11}) - E(\beta_i D_i^{00}). \end{aligned}$$

We have that

$$\begin{aligned} E(\beta_i D_i^{00}) =& \sum_t E(\beta_i D_i^{00}|T = t) \cdot \pi_t \\ =& E(Y_i^1 - Y_i^0|T = at) \cdot \pi_{at} + E(Y_i^1 - Y_i^0|T = 1d) \cdot \pi_{1d} + E(Y_i^1 - Y_i^0|T = 2d) \cdot \pi_{2d} \\ &+ E(Y_i^1 - Y_i^0|T = ed) \cdot \pi_{ed} \end{aligned}$$

and

$$\begin{aligned} E(\beta_i D_i^{11}) =& \sum_t E(\beta_i D_i^{11}|T = t) \cdot \pi_t \\ =& E(Y_i^1 - Y_i^0|T = at) \cdot \pi_{at} + E(Y_i^1 - Y_i^0|T \in cc) \cdot \pi_{cc} + E(Y_i^1 - Y_i^0|T = 1d) \cdot \pi_{1d} \\ &+ E(Y_i^1 - Y_i^0|T = 2d) \cdot \pi_{2d} + E(Y_i^1 - Y_i^0|T = ed) \cdot \pi_{ed}. \end{aligned}$$

Therefore

$$E(Y|Z_1 = 1, Z_2 = 1) - E(Y|Z_1 = 0, Z_2 = 0) = E(\beta_i D_i^{11}) - E(\beta_i D_i^{00}) = E(Y^1 - Y^0|T \in cc) \cdot \pi_{cc},$$

and so

$$\begin{aligned} \beta =& \frac{E\left(Y \mid Z_1 = 1, Z_2 = 1\right) - E\left(Y \mid Z_1 = 0, Z_2 = 0\right)}{E\left(D \mid Z_1 = 1, Z_2 = 1\right) - E\left(D \mid Z_1 = 0, Z_2 = 0\right)} \\ =& \frac{E\left(Y \mid R_2 = 1\right) - E\left(Y \mid R_1 = 1\right)}{E\left(D \mid R_2 = 1\right) - E\left(D \mid R_1 = 1\right)} \\ =& E(Y^1 - Y^0|T \in cc). \end{aligned}$$

B TSLS in subsample with one instrument

Denote the subsample averages of Y and D when $(z_1 = 0, z_2 = 0)$ by $\bar{Y}_{00}$ and $\bar{D}_{00}$, respectively, and as $\bar{Y}_{11}$, and $\bar{D}_{11}$ when $(z_1 = 1, z_2 = 1)$. Denote the total number of observations in the subsample by $\widetilde{N}$, the number of observations for which $(z_1 = 0, z_2 = 0)$ as N_{00}, and the number of observations for which $(z_1 = 1, z_2 = 1)$ as N_{11}. Then $N_{11} = \sum_{i=1}^{\widetilde{N}} \widetilde{Z}$ and $N_{00} = \sum_{i=1}^{\widetilde{N}} (1 - \widetilde{Z})$.

$$
\begin{aligned}
\widetilde{Z}'Y &= \sum_{i=1}^{\widetilde{N}} (\widetilde{Z}_i - \bar{\widetilde{Z}})(y_i - \bar{Y}) \\
&= \sum_{i=1}^{\widetilde{N}} \widetilde{Z}_i (y_i - \bar{Y}) - \bar{\widetilde{Z}} \sum_{i=1}^{\widetilde{N}} (y_i - \bar{Y}) \\
&= \sum_{i=1}^{\widetilde{N}} \widetilde{Z}_i (y_i - \bar{Y}) \\
&= N_{11} \frac{1}{N_{11}} \sum_{i=1}^{\widetilde{N}} \widetilde{Z}_i (y_i - \bar{Y}) \\
&= N_{11} (\bar{y_1} - \bar{Y}) \\
&= N_{11} \left(\bar{y_1} - \frac{N_{00}}{\widetilde{N}} \bar{y_0} - \frac{N_{11}}{\widetilde{N}} \bar{y_1} \right) \\
&= N_{11} \left(\frac{N_{00}\bar{Y}_{11} + N_{11}\bar{Y}_{11}}{\widetilde{N}} - \frac{N_{00}\bar{Y}_{00} + N_{11}\bar{Y}_{11}}{\widetilde{N}} \right) \\
&= \frac{N_{11}N_{00}(\bar{Y}_{11} - \bar{Y}_{00})}{\widetilde{N}}
\end{aligned}
$$

In a similar fashion, one can show that $\widetilde{Z}'D = \frac{N_{11}N_{00}(\bar{D}_{11} - \bar{D}_{00})}{\widetilde{N}}$. Then:

$$
\hat{\beta} = (\widetilde{Z}'D)^{-1}\widetilde{Z}'Y = \frac{N_{11}N_{00}(\bar{Y}_{11} - \bar{Y}_{00})/\widetilde{N}}{N_{11}N_{00}(\bar{D}_{11} - \bar{D}_{00})/\widetilde{N}} = \frac{\bar{Y}_{11} - \bar{Y}_{00}}{\bar{D}_{11} - \bar{D}_{00}}.
$$

C Alternative estimation approaches

Define the following four variables:

$$
R_{1i} = (1 - Z_{1i})(1 - Z_{2i}), \quad R_{2i} = Z_{1i}Z_{2i}, \quad R_{3i} = (1 - Z_{1i})Z_{2i}, \quad R_{4i} = Z_{1i}(1 - Z_{2i}).
$$

A simple consistent estimator of the CC-LATE then consists of the following steps[20]:

1. Use ordinary least squares to estimate the coefficients α_1 and α_2 in

$$
D_i = \alpha_1 R_{1i} + \alpha_2 R_{2i} + \alpha_3 R_{2i} + \alpha_4 R_{2i} + e_i
$$

   where e_i is the regression error. Denote the estimates $\widehat{\alpha}_j$.

2. Use ordinary least squares to estimate the coefficients γ_1 and γ_2 in

$$
Y_i = \gamma_1 R_{1i} + \gamma_2 R_{2i} + \gamma_3 R_{3i} + \gamma_4 R_{4i} + \varepsilon_i
$$

   where ε_i is the regression error. Denote the estimates $\widehat{\gamma}_j$.

[20] As they are unconditionally uncorrelated with R_1 and R_4 by construction, one could drop R_2 and R_3 from these regressions without changing the estimates. However, including them is necessary if one wants to include covariates.

3. The CC-LATE estimator is then

$$\widehat{\beta} = \frac{\widehat{\gamma}_2 - \widehat{\gamma}_1}{\widehat{\alpha}_2 - \widehat{\alpha}_1}.$$

The asymptotic distributions of $\widehat{\beta}$ and $\widehat{\delta}$ can be obtained by the delta method. We can rewrite the above steps as a method of moments (MM) estimator, and use a standard MM estimation package to automatically generate consistent estimates and standard errors. To do so, observe that the above regressions can be expressed as the following set of moments:

$$\begin{aligned} &E\left(\left(D_i - \alpha_1 R_{1i} - (\delta + \alpha_1) R_{2i} - \alpha_3 R_{3i} - \alpha_4 R_{4i}\right) R_{ji}\right) = 0 \quad \text{for } j = 1,2,3,4, \text{ and} \\ &E\left(\left(Y_i - \gamma_1 R_{1i} - (\beta\delta + \gamma_1) R_{2i} - \gamma_3 R_{3i} - \gamma_4 R_{4i}\right) R_{ji}\right) = 0 \quad \text{for } j = 1,2,3,4. \end{aligned} \tag{9}$$

Let the vector $\theta = (\beta, \delta, \alpha_1, \alpha_3, \alpha_4, \gamma_1, \gamma_3, \gamma_4)$. Then, the above eight moments can be replaced with corresponding sample moments, and the parameters θ can be directly estimated using MM estimation. The corresponding $\widehat{\delta}$ will equal $\widehat{\alpha}_2 - \widehat{\alpha}_1$, the estimated probability of an individual i being a combined complier, and $\widehat{\beta}$ will equal the CC-LATE estimate $\frac{\widehat{\gamma}_2 - \widehat{\gamma}_1}{\widehat{\alpha}_2 - \widehat{\alpha}_1}$.

Alternatively, simplifications in getting the limiting distribution of $\widehat{\beta}$ with the delta method can be obtained as follows: Let $\delta = \alpha_2 - \alpha_1$, let $\zeta = \gamma_1 + \gamma_2$, and let $\widetilde{R}_i = R_{1i} + R_{2i}$. Then

$$\begin{aligned} D_i &= \alpha_1 \widetilde{R}_i + \delta R_{2i} + \alpha_3 R_{3i} + \alpha_4 R_{4i} + e_i, \\ Y_i &= \gamma_1 \widetilde{R}_i + \zeta R_{2i} + \gamma_3 R_{3i} + \gamma_4 R_{4i} + \varepsilon_i. \end{aligned}$$

So one can just estimate the OLS regressions of D_i and Y_i on $\widetilde{R}_i$, R_{2i}, R_{3i}, and R_{4i}, and the coefficients of R_{2i} will be consistent estimates of ζ and δ, and $\beta = \zeta/\delta$. Note that we can also set up the MM estimator this way.

D Proof for the extension to multiple instruments

Suppose we have $k > 2$ binary instruments that all satisfy the LATE assumptions. Define $D^{z_1 z_2 \cdots z_k}$ the potential treatment state, $R_1 = (1-Z_1)(1-Z_2)\ldots(1-Z_k)$, and $R_2 = Z_1 Z_2 \ldots Z_k$. Under SUTVA the observed treatment D_i can be written as

$$D_i \;=\; D_i^{00\ldots0} R_{1i} + D_i^{11\ldots1} R_{2i} + \tilde{D}_i$$

where $\tilde{D}_i$ includes all possible combinations of instruments value and the respective potential treatment states.

Thus,

$$\begin{aligned} E\left(D|Z_1 = 1, \ldots, Z_k = 1\right) - E\left(D|Z_1 = 0, \ldots, Z_k = 0\right) &= E(D|R_2 = 1) - E(D|R_1 = 1) \\ &= D_i^{11\ldots1} - D_i^{00\ldots0} \end{aligned}$$

Let cc the set of all complier types, then

$$E\left(D|Z_1=1,\ldots,Z_k=1\right)-E\left(D|Z_1=0,\ldots,Z_k=0\right) \quad = \quad \pi_{cc}.$$

Similarly, it is easy to show that

$$E\left(Y|Z_1=1,\ldots,Z_k=1\right)-E\left(Y|Z_1=0,\ldots,Z_k=0\right) \quad = \quad E(Y^1-Y^0|T\in cc)\pi_{cc}.$$

Thus

$$\frac{E\left(Y|Z_1=1,\ldots,Z_k=1\right)-E\left(Y|Z_1=0,\ldots,Z_k=0\right)}{E\left(D|Z_1=1,\ldots,Z_k=1\right)-E\left(D|Z_1=0,\ldots,Z_k=0\right)}=E(Y^1-Y^0|T\in cc).$$

Another way to obtain this result is as follows:

$$\begin{aligned}
&E\left(Y_i|Z_{1i}=1,\ldots,Z_{ki}=1\right)-E\left(Y|Z_{1i}=0,\ldots,Z_{ki}=0\right)\\
&=E\left(Y_i|R_{2i}=1\right)-E\left(Y_i|R_{1i}=1\right)\\
&=E(D_i^{111\ldots1}\cdot Y_i^1+(1-D_i^{111\ldots1})\cdot Y_i^0|R_{2i}=1)-E(D_i^{000\ldots0}\cdot Y_i^1+(1-D_i^{000\ldots0})\cdot Y_i^0|R_{1i}=1)\\
&=E(D_i^{111\ldots1}\cdot Y_i^1+(1-D_i^{111\ldots1})\cdot Y_i^0)-E(D_i^{000\ldots0}\cdot Y_i^1+(1-D_i^{000\ldots0})\cdot Y_i^0)\\
&=E(D_i^{111\ldots1}\cdot Y_i^1+(1-D_i^{111\ldots1})\cdot Y_i^0-D_i^{000\ldots0}\cdot Y_i^1-(1-D_i^{000\ldots0})\cdot Y_i^0)\\
&=E(D_i^{111\ldots1}\cdot Y_i^1+Y_i^0-D_i^{111\ldots1}\cdot Y_i^0-D_i^{000\ldots0}\cdot Y_i^1-Y_i^0+D_i^{000\ldots0}\cdot Y_i^0)\\
&=E(D_i^{111\ldots1}\cdot Y_i^1-D_i^{111\ldots1}\cdot Y_i^0-D_i^{000\ldots0}\cdot Y_i^1+D_i^{000\ldots0}\cdot Y_i^0)\\
&=E((D_i^{111\ldots1}-D_i^{000\ldots0})(Y_i^1-Y_i^0))\\
&=E\left(E((D_i^{111\ldots1}-D_i^{000\ldots0})(Y_i^1-Y_i^0)|(D_i^{111\ldots1}-D_i^{000\ldots0}))\right)\\
&=1\cdot P(D_i^{111\ldots1}-D_i^{000\ldots0}=1)\cdot E\left(Y_i^1-Y_i^0\middle|\, D_i^{111\ldots1}-D_i^{000\ldots0}=1\right)\\
&\quad -1\cdot P(D_i^{111\ldots1}-D_i^{000\ldots0}=-1)\cdot E\left(Y_i^1-Y_i^0\middle|\, D_i^{111\ldots1}-D_i^{000\ldots0}=-1\right)\\
&\quad +0\cdot P(D_i^{111\ldots1}-D_i^{000\ldots0}=0)\cdot E\left(Y_i^1-Y_i^0|D_i^{111\ldots1}-D_i^{000\ldots0}=0\right)\\
&=E(Y_i^1-Y_i^0|D_i^{111\ldots1}>D_i^{000\ldots0})\cdot P(D_i^{111\ldots1}>D_i^{000\ldots0})\\
&\quad -E(Y_i^1-Y_i^0|D_i^{111\ldots1}<D_i^{000\ldots0})\cdot P(D_i^{111\ldots1}<D_i^{000\ldots0})
\end{aligned}$$

Limited monotonicity rules out the second part (if limited monotonicity is violated then, similar to setting with one binary instrument, treatment effects might be positive for all individuals, but the effect of the defiers cancels out the effect of the compliers). Rewriting leads to the CC-LATE:

$$E\left(Y_i|Z_{1i}=1,\ldots,Z_{ki}=1\right)-E\left(Y|Z_{1i}=0,\ldots,Z_{ki}=0\right)=E(Y_i^1-Y_i^0|D_i^{111\ldots1}>D_i^{000\ldots0})\cdot P(D_i^{111\ldots1}>D_i^{000\ldots0})$$

$$E(Y_i^1 - Y_i^0 | D_i^{111\ldots1} > D_i^{000\ldots0}) = \frac{E\left(Y_i | Z_{1i} = 1, \ldots, Z_{ki} = 1\right) - E\left(Y | Z_{1i} = 0, \ldots, Z_{ki} = 0\right)}{P(D_i^{111\ldots1} > D_i^{000\ldots0})}$$

$$E(Y^1 - Y^0 | T \in cc) = \frac{E\left(Y_i | Z_{1i} = 1, \ldots, Z_{ki} = 1\right) - E\left(Y | Z_{1i} = 0, \ldots, Z_{ki} = 0\right)}{P(D_i^{111\ldots1} - D_i^{000\ldots0} = 1)}$$

$$E(Y^1 - Y^0 | T \in cc) = \frac{E\left(Y_i | Z_{1i} = 1, \ldots, Z_{ki} = 1\right) - E\left(Y | Z_{1i} = 0, \ldots, Z_{ki} = 0\right)}{P(D_i^{111\ldots1} | Z_{1i} = 1, \ldots, Z_{ki} = 1) - P(D_i^{000\ldots0} = 1 | Z_{1i} = 0, \ldots, Z_{ki} = 0)}$$

$$E(Y^1 - Y^0 | T \in cc) = \frac{E\left(Y_i | Z_{1i} = 1, \ldots, Z_{ki} = 1\right) - E\left(Y | Z_{1i} = 0, \ldots, Z_{ki} = 0\right)}{E\left(D | Z_1 = 1, \ldots, Z_k = 1\right) - E\left(D | Z_1 = 0, \ldots, Z_k = 0\right)}$$

$0 \cdot P(D_i^{111\ldots1} - D_i^{000\ldots0} = 0) \cdot E\left(Y_i^1 - Y_i^0 | D_i^{111\ldots1} - D_i^{000\ldots0} = 0\right)$ demonstrates the fact that the CC-LATE does not capture the effect for those individuals for whom a change from being exposed to none of the instruments to being exposed to all instruments simultaneously does not change the treatment status meaning that this change is not informative for these individuals. The always-takers and never-takers belong to this group.

E CC-LATE under IAM

In their Appendix, Imbens and Angrist (1994) state that under IAM:

$$E(Y|Z = z_K) = E(Y|Z = z_0) + \alpha_{z_K, z_0} \cdot (P(z_K) - P(z_0))$$

We can rewrite this as follows:

$$\frac{E(Y|Z = z_K) - E(Y|Z = z_0)}{P(z_K) - P(z_0)} = \alpha_{z_K, z_0}$$

$$\Downarrow$$

$$\frac{E(Y|Z = z_K) - E(Y|Z = z_0)}{E(D|Z = z_K) - E(D|Z = z_0)} = E(Y(1) - Y(0)|D(z_K) \neq D(z_0))$$

$$\Downarrow$$

$$\frac{E(Y|Z = z_K) - E(Y|Z = z_0)}{E(D|Z = z_K) - E(D|Z = z_0)} = \frac{\sum_{l=1}^{K} \alpha_{z_l, z_{l-1}} \cdot (P(z_l) - P(z_{l-1}))}{P(z_K) - P(z_0)}$$

$$\Downarrow$$

$$\frac{E(Y|Z = z_K) - E(Y|Z = z_0)}{E(D|Z = z_K) - E(D|Z = z_0)} = \sum_{l=1}^{K} \frac{P(z_l) - P(z_{l-1})}{P(z_K) - P(z_0)} \cdot \alpha_{z_l, z_{l-1}}$$

$\frac{E(Y|Z=z_K) - E(Y|Z=z_0)}{E(D|Z=z_K) - E(D|Z=z_0)} = E(Y(1) - Y(0)|D(z_K) \neq D(z_0))$ shows that this can be interpreted as the effect in the largest group of compliers. This is the same interpretation as the estimand for multiple binary instruments as proposed by Frölich (2007).

Suppose we have two binary instruments and the support $z_0 = (0,0),\ z_1 = (0,1),\ z_2 = (1,0),\ z_3 = (1,1)$, ordered such that $l < m$ implies $P_l < P_m$. Then the final line in the last expression can be re-written as:

$$
\begin{aligned}
\alpha_{30} =& \frac{(P_{z_1} - P_{z_0}) \cdot \alpha_{z_1 z_0} + (P_{z_2} - P_{z_1}) \cdot \alpha_{z_2 z_1} + (P_{z_3} - P_{z_2}) \cdot \alpha_{z_3 z_2}}{P_{z_3} - P_{z_0}} \\
=& \frac{(P_{z_1} - P_{z_0})}{P_{z_3} - P_{z_0}} \cdot \frac{E(Y|Z = z_1) - E(Y|Z = z_0)}{P_{z_1} - P_{z_0}} + \frac{(P_{z_2} - P_{z_1})}{P_{z_3} - P_{z_0}} \cdot \frac{E(Y|Z = z_2) - E(Y|Z = z_1)}{P_{z_2} - P_{z_1}} \\
& + \frac{(P_{z_3} - P_{z_2})}{P_{z_3} - P_{z_0}} \cdot \frac{E(Y|Z = z_3) - E(Y|Z = z_2)}{P_{z_3} - P_{z_2}} \\
=& \frac{E(Y|Z = z_1) - E(Y|Z = z_0) + E(Y|Z = z_2) - E(Y|Z = z_1) + E(Y|Z = z_3) - E(Y|Z = z_2)}{P_{z_3} - P_{z_0}} \\
=& \frac{E(Y|Z = z_3) - E(Y|Z = z_0)}{P_{z_3} - P_{z_0}} \\
=& \frac{E(Y|Z = z_3) - E(Y|Z = z_0)}{E(D|Z = z_3) - E(D|Z = z_0)}
\end{aligned}
$$

F Proof for violation of limited monotonicity

Consider the setting where limited monotonicity is violated. Let $\pi_t = \Pr(T \in t)$, $t = at, rc, ec, 1c, 2c, 1d, 2d, ed, rd, d1, d2, d3, d4, d5, d6, nt$. We have

$$
\begin{aligned}
E(D_i^{00}) =& \sum_t E(D_i^{00}|T = t)\pi_t \\
=& \pi_{at} \cdot 1 + \pi_{rc} \cdot 0 + \pi_{ec} \cdot 0 + \pi_{1c} \cdot 0 + \pi_{2c} \cdot 0 + \pi_{nt} \cdot 0 + \pi_{1d} \cdot 1 + \pi_{2d} \cdot 1 + \pi_{ed} \cdot 1 + \pi_{d3} \cdot 1 \\
& + \pi_{d1} \cdot 0 + + \pi_{d4} \cdot 1 + \pi_{d2} \cdot 0 + \pi_{d5} \cdot 1 + \pi_{rd} \cdot 0 + \pi_{d6} \cdot 1 \\
=& \pi_{at} + \pi_{1d} + \pi_{2d} + \pi_{ed} + \pi_{d3} + \pi_{d4} + \pi_{d5} + \pi_{d6}
\end{aligned}
$$

and

$$
\begin{aligned}
E(D_i^{11}) =& \sum_t E(D_i^{11}|T = t)\pi_t \\
=& \pi_{at} \cdot 1 + \pi_{rc} \cdot 1 + \pi_{ec} \cdot 1 + \pi_{1c} \cdot 1 + \pi_{2c} \cdot 1 + \pi_{nt} \cdot 0 + \pi_{1d} \cdot 1 + \pi_{2d} \cdot 1 + \pi_{ed} \cdot 1 + \pi_{d1} \cdot 0 \\
& + \pi_{d2} \cdot 0 + \pi_{rd} \cdot 0 + \pi_{d3} \cdot 0 + \pi_{d4} \cdot 0 + \pi_{d5} \cdot 0 + \pi_{d6} \cdot 0 \\
=& \pi_{at} + \underbrace{\pi_{rc} + \pi_{ec} + \pi_{1c} + \pi_{2c}}_{\pi_{cc}} + \pi_{1d} + \pi_{2d} + \pi_{ed} \\
=& \pi_{at} + \pi_{cc} + \pi_{1d} + \pi_{2d} + \pi_{ed}.
\end{aligned}
$$

It therefore follows that

$$
E(D|Z_1 = 1, Z_2 = 1) - E(D|Z_1 = 0, Z_2 = 0) = E(D_i^{11}) - E(D_i^{00}) = \pi_{cc} - (\pi_{d3} + \pi_{d4} + \pi_{d5} + \pi_{d6}).
$$

Let $\beta_i = Y_i^1 - Y_i^0$. Under SUTVA the observed outcome Y can be written as

$$
\begin{aligned}
Y_i &= Y_i^1 D_i + Y_i^0 (1 - D_i) = \beta_i D_i + Y_i^0 \\
&= \beta_i \left[D_i^{00} R_{1i} + D_i^{11} R_{2i} + D_i^{01} R_{3i} + D_i^{10} R_{4i} \right] + Y_i^0 \\
&= \beta_i D_i^{00} R_{1i} + \beta_i D_i^{11} R_{2i} + \beta_i D_i^{01} R_{3i} + \beta_i D_i^{10} R_{4i} + Y_i^0.
\end{aligned}
$$

Now, consider the numerator of the CC-LATE estimand,

$$\begin{aligned} E\left(Y|Z_1=1,Z_2=1\right)-E\left(Y|Z_1=0,Z_2=0\right) &= E(Y|R_2=1)-E(Y|R_1=1) \\ &= E(\beta_i D_i^{11}+Y_i^0|R_2=1)-E(\beta_i D_i^{00}+Y_i^0|R_1=1) \\ &= E(\beta_i D_i^{11})-E(\beta_i D_i^{00}). \end{aligned}$$

We have that

$$\begin{aligned} E(\beta_i D_i^{00}) =& \sum_t E(\beta_i D_i^{00}|T=t)\cdot\pi_t \\ =& E(Y_i^1-Y_i^0|T=at)\cdot\pi_{at}+E(Y_i^1-Y_i^0|T=1d)\cdot\pi_{1d}+E(Y_i^1-Y_i^0|T=2d)\cdot\pi_{2d} \\ &+E(Y_i^1-Y_i^0|T=ed)\cdot\pi_{ed}+E(Y_i^1-Y_i^0|T=d3)\cdot\pi_{d3}+E(Y_i^1-Y_i^0|T=d4)\cdot\pi_{d4} \\ &+E(Y_i^1-Y_i^0|T=d5)\cdot\pi_{d5}+E(Y_i^1-Y_i^0|T=d6)\cdot\pi_{d6} \end{aligned}$$

and

$$\begin{aligned} E(\beta_i D_i^{11}) =& \sum_t E(\beta_i D_i^{11}|T=t)\cdot\pi_t \\ =& E(Y_i^1-Y_i^0|T=at)\cdot\pi_{at}+E(Y_i^1-Y_i^0|T\in cc)\cdot\pi_{cc}+E(Y_i^1-Y_i^0|T=1d)\cdot\pi_{1d} \\ &+E(Y_i^1-Y_i^0|T=2d)\cdot\pi_{2d}+E(Y_i^1-Y_i^0|T=ed)\cdot\pi_{ed}. \end{aligned}$$

Therefore

$$\begin{aligned} & E(Y|Z_1=1,Z_2=1)-E(Y|Z_1=0,Z_2=0) \\ &= E(\beta_i D_i^{11})-E(\beta_i D_i^{00}) \\ &= E(Y^1-Y^0|T\in cc)\cdot\pi_{cc}-E(Y_i^1-Y_i^0|T=d3)\cdot\pi_{d3} \\ &\quad -E(Y_i^1-Y_i^0|T=d4)\cdot\pi_{d4}-E(Y_i^1-Y_i^0|T=d5)\cdot\pi_{d5} \\ &\quad -E(Y_i^1-Y_i^0|T=d6)\cdot\pi_{d6} \\ =& E(Y^1-Y^0|T\in cc)\cdot\pi_{cc}-E(Y^1-Y^0|T\in dd)\cdot\pi_{dd} \end{aligned}$$

with dd the set of defiers that can never be pushed towards compliance and do not cancel out, $dd\equiv\{d3,d4,d5,d6\}$, and so

$$\begin{aligned} \beta =& \frac{E\left(Y\mid Z_1=1,Z_2=1\right)-E\left(Y\mid Z_1=0,Z_2=0\right)}{E\left(D\mid Z_1=1,Z_2=1\right)-E\left(D\mid Z_1=0,Z_2=0\right)} \\ =& \frac{E(Y^1-Y^0|T\in cc)\cdot\pi_{cc}-E(Y^1-Y^0|T\in dd)\cdot\pi_{dd}}{\pi_{cc}-\pi_{dd}} \\ =& \frac{\pi_{cc}}{\pi_{cc}-\pi_{dd}}E(Y^1-Y^0|T\in cc)-\frac{\pi_{dd}}{\pi_{cc}-\pi_{dd}}E(Y^1-Y^0|T\in dd). \end{aligned}$$

G Proof of Bloom result for multiple instruments

When there are two binary instruments, one-sided compliance means that

$$E(D_i|Z_1=0,Z_2=0)=P(D_i=1|Z_{1i}=0,Z_{2i}=0)=0.$$

We can re-write $E(Y_i|Z_{1i}=1,Z_{2i}=1)$ and $E(Y_i|Z_{1i}=0,Z_{2i}=0)$ as

$$E(Y_i|Z_{1i}=1,Z_{2i}=1)=E(Y_i^0|Z_{1i}=1,Z_{2i}=1)+E((Y_i^1-Y_i^0)D_i|Z_{1i}=1,Z_{2i}=1) \quad (10)$$

and

$$E(Y_i|Z_{1i}=0,Z_{2i}=0)=E(Y_i^0|Z_{1i}=0,Z_{2i}=0)+E((Y_i^1-Y_i^0)D_i|Z_{1i}=0,Z_{2i}=0) \quad (11)$$

where $E((Y_i^1-Y_i^0)D_i|Z_{1i}=0,Z_{2i}=0)=0$ because $D_i=0$ if $Z_{1i}=0,Z_{2i}=0$. Subtracting equation (11) from equation (10) gives

$$\begin{aligned}
&E(Y_i|Z_{1i}=1,Z_{2i}=1)-E(Y_i|Z_{1i}=0,Z_{2i}=0)\\
&=E((Y_i^1-Y_i^0)D_i|Z_{1i}=1,Z_{2i}=1)\\
&=E(Y_i^1-Y_i^0|D_i=1,Z_{1i}=1,Z_{2i}=1)P(D_i=1|Z_{1i}=1,Z_{2i}=1)
\end{aligned}$$

where the first equality follows because $E(Y_i^0|Z_{1i}=1,Z_{2i}=1)=E(Y_i^0|Z_{1i}=0,Z_{2i}=0)$ by the independence assumption.

Note that unlike in the setting with one binary instrument where $D_i=1$ implies $Z_i=1$, in the setting with two binary instruments $D_i=1$ does **not** imply $Z_{1i}=1,Z_{2i}=1$. So $E(Y_i^1-Y_i^0|D_i=1,Z_{1i}=1,Z_{2i}=1)\neq E(Y_i^1-Y_i^0|D_i=1)$. However, if compliance is only possible when both instruments are offered such that $Z_{1i}=1,Z_{2i}=1$, then $E(Y_i^1-Y_i^0|D_i=1,Z_{1i}=1,Z_{2i}=1)=E(Y_i^1-Y_i^0|D_i=1)$, and the treatment effect on the treated is

$$E(Y_i^1-Y_i^0|D_i=1)=\frac{E(Y_i|Z_{1i}=1,Z_{2i}=1)-E(Y_i|Z_{1i}=0,Z_{2i}=0)}{P(D_i=1|Z_{1i}=1,Z_{2i}=1)}.$$

This result can easily be extended to the setting with more than two binary instruments if it holds that compliance is only possible when an individual is exposed to all instruments.

H Tables with application estimates

Table 11: Estimates corresponding to Figure 3a.

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Bought condoms						
estimates	0.024	0.288	0.228	1.854	0.17	0.011
(std. err.)	(0.033)	(0.157)	(0.139)	(6.424)	(0.116)	(0.093)
nr. obs.	1008	1008	1008	1008	1008	1008
Panel B: Number of condoms bought						
estimates	-0.035	0.94	0.799	5.906	0.521	-0.199
(std. err.)	(0.139)	(0.489)	(0.662)	(144.096)	(0.404)	(0.4)
nr. obs.	1008	1008	1008	1008	1008	1008
Panel C: Reported buying condoms						
estimates	-0.009	0.096	0.161	2.07	-0.022	0.051
(std. err.)	(0.025)	(0.087)	(0.069)	(2.965)	(0.06)	(0.046)
nr. obs.	1008	1008	1008	1008	1008	1008
Panel D: Reported having sex						
estimates	0.032	0.023	0.022	0.22	0.019	0.054
(std. err.)	(0.033)	(0.146)	(0.114)	(20.002)	(0.063)	(0.116)
nr. obs.	1008	1008	1008	1008	1008	1008

The columns give the estimates for the different methods: (1) $\hat{\beta}_{OLS}$, (2) $\hat{\beta}_{CC-LATE-2}$, (3) $\hat{\beta}_{CC-LATE-3}$, (4) $\hat{\beta}_{TSLS-above-1.5km-distance}$, (5) $\hat{\beta}_{TSLS-any-cash}$, and (6) $\hat{\beta}_{TSLS-above-median-cash}$.

Cluster bootstrapped standard errors with 1000 repetitions.

Table 12: Estimates corresponding to Figure 3b.

	$\hat{\beta}_{CC-LATE-2}$	$\hat{\beta}_{CC-LATE-3}$	$\hat{\beta}_{TSLS-2}$	$\hat{\beta}_{TSLS-3}$
Panel A: Bought condoms				
estimates	0.288	0.228	0.177	0.118
(std. err.)	(0.157)	(0.139)	(0.135)	(0.106)
nr. obs.	1008	1008	1008	1008
Panel B: Number of condoms bought				
estimates	0.94	0.799	0.543	0.278
(std. err.)	(0.489)	(0.662)	(0.478)	(0.337)
nr. obs.	1008	1008	1008	1008
Panel C: Reported buying condoms				
estimates	0.096	0.161	-0.013	0.012
(std. err.)	(0.087)	(0.069)	(0.044)	(0.052)
nr. obs.	1008	1008	1008	1008
Panel D: Reported having sex				
estimates	0.023	0.022	0.02	0.032
(std. err.)	(0.146)	(0.114)	(0.095)	(0.058)
nr. obs.	1008	1008	1008	1008

Cluster bootstrapped standard errors with 1000 repetitions.

I Figure 3a without the distance instrument

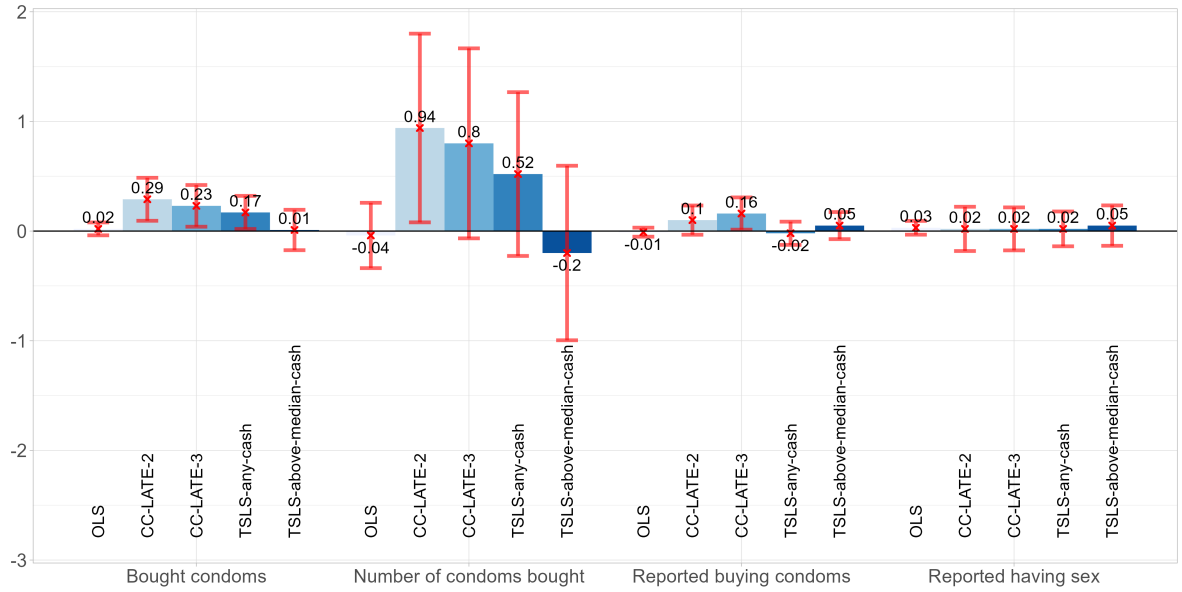


Figure 4: Figure 3a without the distance instrument to allow for easier comparison of the CC-LATE estimator with the LATEs of each instrument used separately.

J Simulation study

In this section, we perform two different simulation studies on the performance of the CC-LATE estimator. First, we compare the CC-LATE estimator to the TSLS estimator when PM is valid and when PM is violated. Second, we compare the performance of the CC-LATE estimator when adding a weak third instrument to adding a strong third instrument.

J.1 Comparison of the CC-LATE to the TSLS estimator when PM is violated

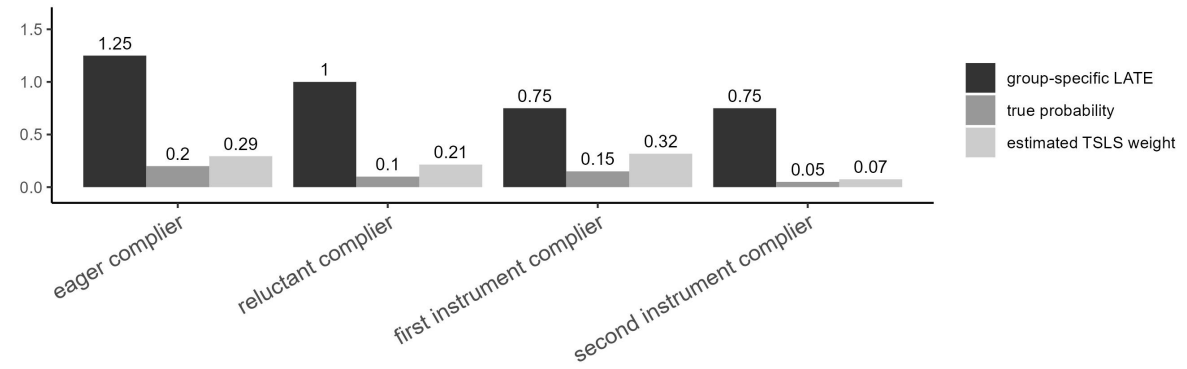
J.1.1 Setup

Following the idea of an Empirical Monte Carlo Study (EMCS) by Huber et al. (2013), the DGP of the simulation study largely depends on the real data of the HIV application studied in Section 6. We investigate the performance of the CC-LATE estimator compared to the TSLS estimator in two different settings. In the first setting, PM is valid. In the second setting, PM is violated due to the presence of defier types. Potential threats in the HIV application are the existence of second instrument defiers or defiers of type 1. This could lead to a violation of PM while LiM would still hold.

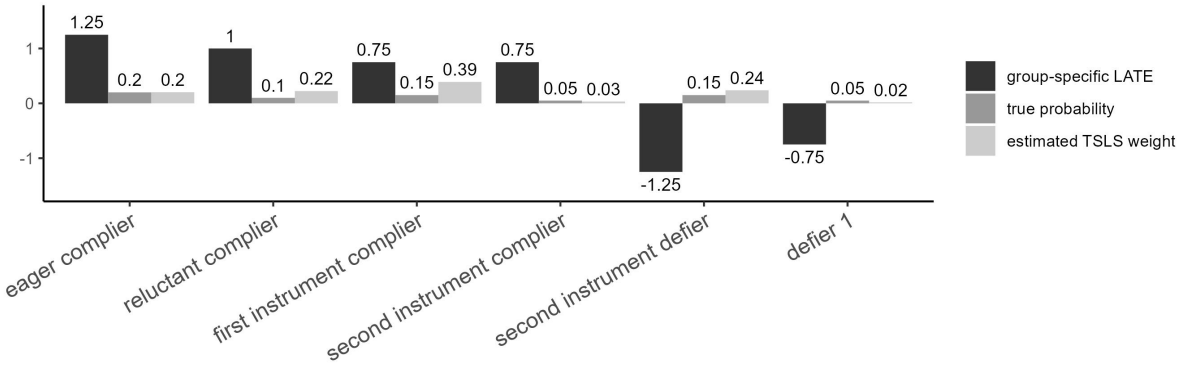
Figure 5 depicts the true probabilities and the average effects per response type used in the simulation[21]. In Section 6, the estimated CC-LATE for the number of condoms bought when using two instruments is 0.8 and we use similar values for choosing the group-specific LATEs, β_{t_i}, of each response type. The probabilities of belonging to a certain response type are chosen based on the information that can be obtained from the HIV application. Under LiM, the response group proportions $\pi_{rd} + \pi_{d1} + \pi_{d2} + \pi_{nt}$ and $\pi_{at} + \pi_{1d} + \pi_{2d} + \pi_{ed}$ can be estimated (see Table 7 in Section 8). Under PM, the defier types are ruled out such that π_{nt} and π_{at} can be estimated. We estimate these probabilities for the HIV application. We further use the estimated shares of the complier population from Figure 2 in Section 6. With these group-specific LATEs and pre-defined probabilities, the true value of the LATE for the combined compliers equals 1.

The sample size is $n = 1000$ which is similar to the 1008 observations of the HIV application. The instruments, Z_1 and Z_2, are drawn from a Bernouilli distribution with the probability set to the mean of the two binary instruments from the application, *any cash* and *distance*. Similar to the application where the instruments are randomized, the instruments are independent of each other. The response types, t_i, are sampled with the pre-defined probabilities. The value of D_i is then set based on the sampled response type and the instrument values. In the sample of

[21] Figure 5 also contains the estimated TSLS weights using equation (20) and (21) from the proof of Proposition 7 in Mogstad et al. (2021). To calculate the weights, propensity scores are predicted nonparametrically. The weights do not exactly add up to one since they are estimated. The weights are nonnegative since our simulation study considers the setting where the instruments are monotonic in the propensity score which is the most realistic scenario considering the HIV application.



(a) True LATEs, true weights and estimated TSLS weights when PM is not violated.



(b) True LATEs, true weights and estimated TSLS weights when PM is violated.

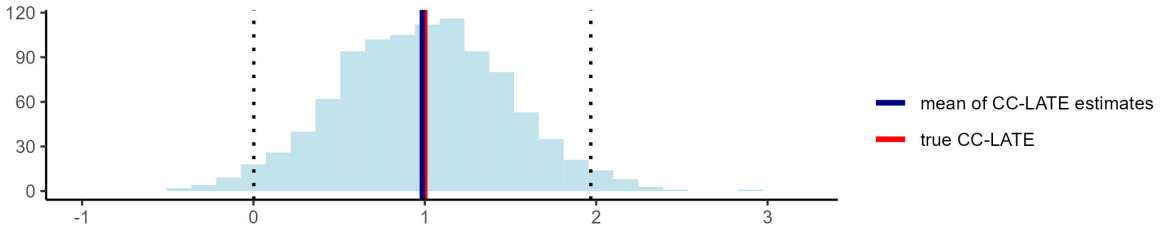
Figure 5: This figure contains the true LATEs and true weights used in the simulation study. It further shows the estimated TSLS weights when PM holds compared to when it is violated.

untreated individuals, we calculate the mean, m_y, and the variance, v_y, of the outcome on the number of condoms bought. Then, $Y_i(0) = m_y$ and $Y_i = m_y + \beta_{t_i} D_i + \nu_i$ where $\nu_i \sim N(0, v_y)$. We perform 1000 simulation repetitions.

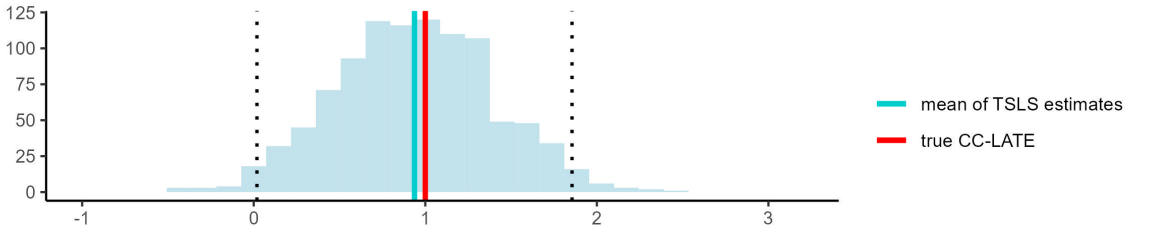
J.1.2 Results

We compare the performance of the CC-LATE estimator and the TSLS estimator when PM is violated due to the presence of defier types. The estimates are compared to the true value of the LATE for the combined compliers, assuming that the subjective of both methods is to give an estimate of the ATE for this subpopulation. Note that this objective is true for TSLS if PM is imposed such that increasing the instrument values weakly increases treatment uptake, as in Section 6.

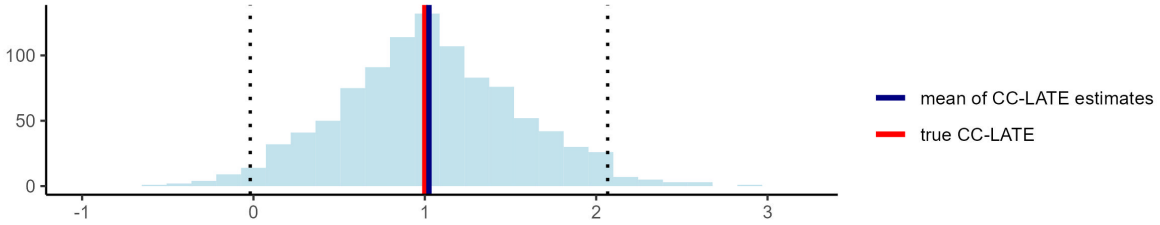
The distributions of the estimates are depicted in Figure 6, and Table 13 gives the bias, median bias, mean absolute error (MAE), and mean squared error (MSE). The MSE and MAE of the CC-LATE and TSLS estimator are comparable, since the CC-LATE estimates lie closer



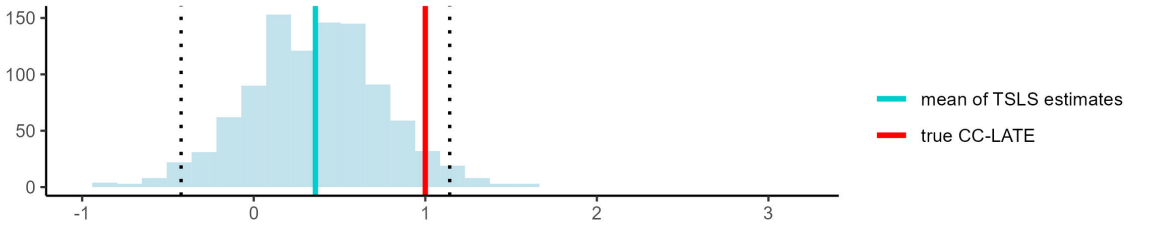
(a) Distribution of CC-LATE estimates when PM is valid.



(b) Distribution of TSLS estimates when PM is valid.



(c) Distribution of CC-LATE estimates when PM is violated.



(d) Distribution of TSLS estimates when PM is violated.

Figure 6: This figure compares the distributions of CC-LATE and TSLS estimates when PM is valid versus when PM is violated. 95% confidence intervals indicated by dashed lines.

to the true value but are more spread out than the TSLS estimates. When PM holds, both the CC-LATE estimator and the TSLS estimator lie close to the true LATE for the combined compliers. Even though the CC-LATE estimator uses fewer observations, the standard deviation of the estimates of the two methods is comparable. Violation of PM clearly introduces downward bias in the TSLS estimates since it now includes the LATEs of the second instrument defiers and the defiers of type 1. Interestingly, this might also explain the smaller coefficients found with TSLS in Section 6. Interestingly, this provides some informal evidence in favor of the existence of defier types in the HIV application. As LiM still holds in the presence of the introduced defier types, the bias of the CC-LATE estimator remains small when PM is violated.

Table 13: This table contains the estimates and measures compared to the true LATE for the combined complier population when PM is valid and when PM is violated.

	(1)		**(2)**	
	PM valid		**PM violated**	
	CC-LATE estimator	TSLS estimator	CC-LATE estimator	TSLS estimator
mean of estimates	0.985	0.936	1.024	0.363
std. dev. of estimates	0.492	0.467	0.521	0.392
bias	-0.015	-0.064	0.024	-0.637
median bias	-0.021	-0.068	0.001	-0.624
MSE	0.242	0.222	0.272	0.560
MAE	0.395	0.376	0.409	0.656

J.2 Comparing CC-LATE estimators when adding a third (weak) instrument

J.2.1 Setup

In this section, we study the performance of the CC-LATE estimator for two different settings where a third instrument is available. The DGPs are similar to the DGPs in Section J.1. In the first setting, the third instrument is extremely weak in that it pushes none of the individuals to compliance. The third instrument, Z_3, is drawn from a Bernouilli distribution with the probability equal to the mean of the *above median cash* instrument from the HIV application. The types considered in this simulation study are given in Table 14. The response types are chosen such that there are only compliers with respect to Z_1 and Z_2. Using similar notions as in the setting with two instruments, these are the eager compliers, reluctant compliers, first instrument compliers, and second instrument compliers with respect to Z_1 and Z_2. In the second setting, the third instrument is strong and adds compliers that only respond to this instrument. The third instrument complier type always takes up treatment when exposed to

the third instrument, but does not influence the complier population when exposed to Z_1 or Z_2, since these response types are either always-takers or never-takers when Z_3 is fixed. The table with all probabilities and group-specific LATEs used in simulation can be found in Table 15. For the second setting, the probability of being a third instrument complier equals 20%. So in this setting, the third instrument pushes many individuals towards compliance.

Table 14: Table with types considered in the simulation study.

D^{111}	D^{110}	D^{101}	D^{011}	D^{100}	D^{010}	D^{001}	D^{000}	Type when $Z_3 = 1$	Type when $Z_3 = 0$	Notion
1	1	1	1	1	1	1	1	always-taker	always-taker	always-taker
1	1	1	1	1	1	0	0	eager complier	eager complier	eager complier
1	1	0	0	0	0	0	0	reluctant complier	reluctant complier	reluctant complier
1	1	1	0	1	0	0	0	first instrument complier	first instrument complier	first instrument complier
1	1	0	1	0	1	0	0	second instrument complier	second instrument complier	second instrument complier
1	0	1	1	0	0	1	0	always-taker	never-taker	third instrument complier
0	0	0	0	0	0	0	0	never-taker	never-taker	never-taker

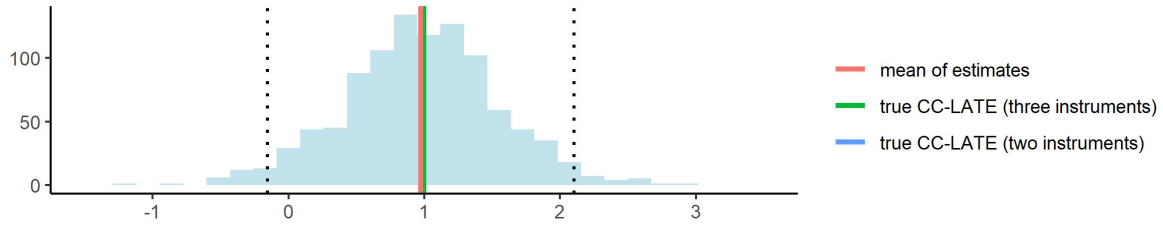
Table 15: Table with true average treatment effects and probabilities per response type. We compare the setting where the third instrument does not add compliers to the setting where it adds compliers.

	(1)		**(2)**	
	Third instrument does not add compliers		**Third instrument adds compliers**	
Response type	**Probability**	**True LATE**	**Probability**	**True LATE**
always-taker	0.4	0	0.3	0
eager complier	0.2	1.25	0.2	1.25
reluctant complier	0.05	0.5	0.05	0.5
first instrument complier	0.15	1	0.15	1
second instrument complier	0.05	0.5	0.05	0.5
third instrument complier			0.2	1.5
never-taker	0.15	0	0.05	0
true CC-LATE two inst.		1		1
true CC-LATE three inst.		1		1.154

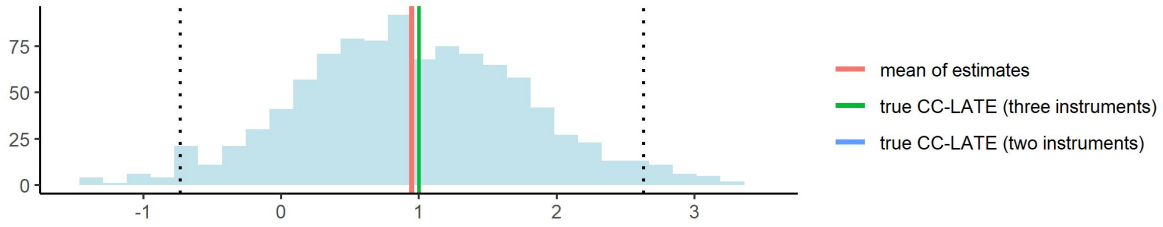
J.2.2 Results

We estimate the CC-LATE using either two or three instruments where the third instrument is either weak or strong. Figure 7 depicts the estimate distributions[22]. When including a third instrument that does not add any compliers, the estimated CC-LATE lies close to the true LATE of the combined compliers which consist of the eager compliers, reluctant compliers, first instrument compliers, and second instrument compliers in this case (see Figure 7a). Since adding a third instrument reduces the number of observations used for estimation, the confidence

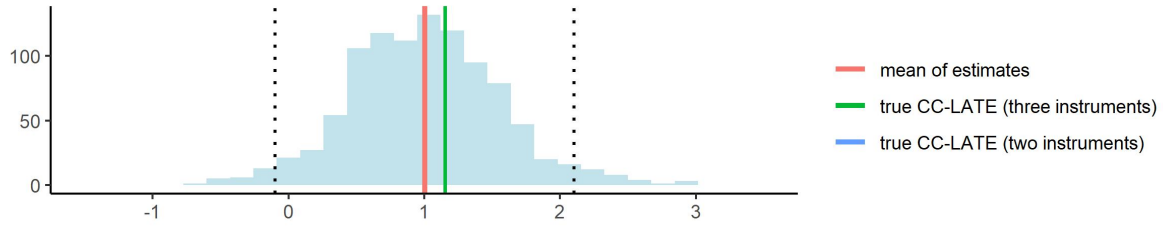
[22]Table 16 contains mean of estimate and standard deviations corresponding to Figure 7.



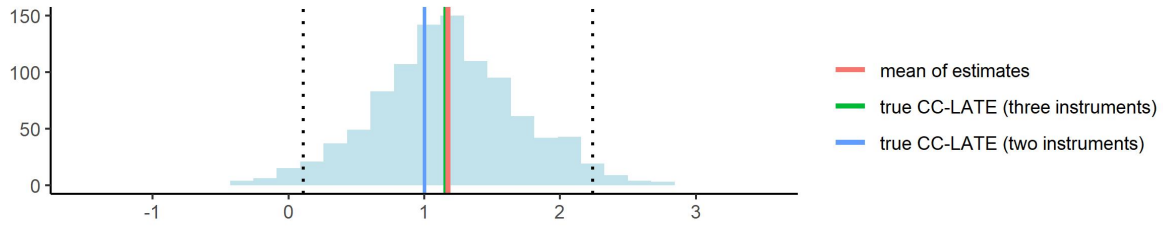
(a) Distribution of the CC-LATE estimates when using two instruments.



(b) Distribution of the CC-LATE estimates when using three instruments where the third instrument does not add any compliers.



(c) Distribution of the CC-LATE estimates when using two instruments and leaving out the third instrument while there are third instrument compliers present in the population.



(d) Distribution of the CC-LATE estimates when using three instruments where the third instrument adds third instrument compliers to the complier population.

Figure 7: This figure compares the distributions of the CC-LATE estimates for settings where two or three instruments are used where the third instrument either adds to the complier population or does not add any compliers at all. 95% confidence intervals indicated by dashed lines.

Table 16: Table with CC-LATE estimates of in case of two or three binary instruments for the settings where the third instrument does add third instrument compliers and where it does not add compliers, corresponding to Figure 7.

	(1)		**(2)**	
	Third instrument does not add compliers		**Third instrument adds compliers**	
	two instruments	three instruments	two instruments	three instruments
mean of estimates	0.975	0.949	1.003	1.175
std. dev. of estimates	0.565	0.842	0.551	0.532
bias	-0.025	-0.051	0.003	0.021
median bias	-0.021	-0.085	0.005	0.016
MSE	0.319	0.710	0.303	0.284
MAE	0.442	0.671	0.432	0.417

intervals are wider (see Figure 7b).

When third instrument compliers are present in the population, the mean of the CC-LATE estimates using only two instruments, Z_1 and Z_2, lies close to the true CC-LATE for the combined complier population with respect to these two instruments (see Figure 7c). Including a strong third instrument that adds third instrument compliers leads to an increase in the complier population considered. It estimates the LATE for the eager compliers, reluctant compliers, first instrument compliers, and second instrument compliers as well as the third instrument compliers (see Figure 7d).

Concluding, when an extremely weak instrument is added, the CC-LATE remains unbiased but is less precise. When incorporating the additional instrument, the compliers that respond to this instrument are added to the complier population. While, the precision remains approximately the same, the estimated LATE considers a larger subpopulation and hence might lie closer to the true ATE.