Interacting Treatments With Endogenous Takeup

Kormos, Máté; Lieli, Robert P.; Huber, Martin

DOI
10.1002/jae.3120

Publication date
2025

Document Version
Final published version

Published in
Journal of Applied Econometrics

Citation (APA)
Kormos, M., Lieli, R. P., & Huber, M. (2025). Interacting Treatments With Endogenous Takeup. *Journal of Applied Econometrics*, *40*(4), 424-437. https://doi.org/10.1002/jae.3120

Journal of Applied Econometrics

WILEY

Journal of APPLIED ECONOMETRICS

RESEARCH ARTICLE

Interacting Treatments With Endogenous Takeup

Máté Kormos[1] | Robert P. Lieli[2] | Martin Huber[3]

[1]Delft University of Technology, Delft, Netherlands | [2]Central European University, Vienna, Austria | [3]University of Fribourg, Fribourg, Switzerland

Correspondence: Robert P. Lieli (lielir@ceu.edu)

Received: 11 June 2024 | **Revised:** 21 November 2024 | **Accepted:** 13 January 2025

Keywords: causal inference | interaction | instrumental variables | non-compliance

ABSTRACT

We study causal inference in randomized experiments (or quasi-experiments) following a 2×2 factorial design. There are two treatments, denoted A and B, and units are randomly assigned to one of four categories: treatment A alone, treatment B alone, joint treatment, or none. Allowing for endogenous non-compliance with the two binary instruments representing the intended assignment, as well as unrestricted interference across the two treatments, we derive the causal interpretation of various instrumental variable estimands under more general compliance conditions than in the literature. In general, if treatment takeup is driven by both instruments for some units, it becomes difficult to separate treatment interaction from treatment effect heterogeneity. We provide auxiliary conditions and various bounding strategies that may help zero in on causally interesting parameters. We apply our results to a program randomly offering two different treatments to first-year college students, namely, tutoring and financial incentives, in order to assess the effect of the treatments on academic performance.

JEL Classification: C22, C26, C90

1 | Introduction

The experimental approach to establishing the causal effect of a treatment is based on allocating units randomly to treatment and control, thereby precluding any systematic difference between the two groups other than the treatment itself. While unit-level treatment effects may be heterogeneous, comparing the average outcome in the treated and control groups gives a consistent estimate of the average treatment effect. This deceptively simple description of the experimental ideal, which originates in the work of Fisher (1925), embodies several further assumptions, formalized later by Rubin (1974, 1978) and others. A lot of subsequent work on causal inference has sought to extend the analysis of experimental data to more complicated situations with some of the following features.

First, in real-world experiments, perfect compliance with the intended treatment assignment is not always possible or ethical to enforce. If non-compliance is endogenous (i.e., it depends on unobserved confounders), then the average difference between the treated and non-treated values does not represent the treatment effect alone but selection effects as well. Second, randomization itself does not ensure that the treatment status of an individual does not interfere with the potential outcomes of another, violating what is called the stable unit treatment value assumption (SUTVA) in the Rubin causal model. Third, there are experimental setups in which units have access to multiple, but not mutually exclusive, treatments that may interact with each other.

In this paper, we derive the causal interpretation of various instrumental variable (IV) estimands in an experimental or quasi-experimental setup that extends the basic model in all three directions mentioned above. More concretely:

i. Population units are targeted by two binary (0/1) treatments, denoted as D_A and D_B.

ii. The individual treatment effects and the extent to which the two treatments interact may vary from unit to unit in an arbitrary way.

iii. There are two binary (0/1) instruments, denoted as Z_A and Z_B, representing randomized assignment to the corresponding treatment or some exogenous incentive to take that treatment. Nevertheless, compliance is imperfect (endogenous), including the possibility of "instrument spillovers," where Z_A affects the takeup of D_B or Z_B affects the takeup of D_A. We also refer to this situation as "coordinated compliance" (see Example 2).

We subsequently present two examples, which illustrate these ideas.

Example 1. Angrist, Lang, and Oreopoulos (2009) assess a randomized program providing two treatments to first-year college students: academic services in the form of tutoring (D_A) and financial incentives (D_B), both aimed at improving academic performance. Students who entered a Canadian college in September 2005 and had a high school grade point average lower than the upper quartile were randomly offered either one, both, or no treatment. Therefore, the instruments Z_A and Z_B are binary indicators for being offered tutoring and/or financial incentives. While the offers are randomly assigned, the actual treatment takeup is likely to be endogenous as it might be driven by personality traits also affecting academic performance. In addition, the treatments may interact such that, for instance, the effectiveness of financial incentives might increase when also receiving tutoring. Section 5 provides an illustration of our results based on this example.

Example 2. The population of interest comprises of married couples where one member suffers from depression and the other does not. There are two binary treatments: an antidepressant medication for the depressed spouse (D_A), and an educational program about depression for the healthy spouse (D_B). The dependent variable may measure the severity of the depression symptoms. Even if the intended treatment assignments (Z_A and Z_B) are random, the actual compliance decision may well be endogenous and coordinated across the spouses (for example, they might agree that they will take the treatments if and only if both of them are assigned). Moreover, the two treatments may interact; the medication might be more effective if accompanied by behavioral adjustments on the partner's part.

Motivated by Example 2, it will be convenient in the rest of the paper to represent population units as pairs (A, B), where member A is targeted by D_A/Z_A and member B is targeted by D_B/Z_B. The outcome of interest may be associated with member A, member B or the pair itself. The representation of units in terms of pairs comes without loss of generality, as a single individual targeted by two treatments (such as in Example 1) can always be thought of as a "pair" with identical members. Thus, we can equate treatment interaction with interference across pair members (a violation of SUTVA).

The estimands we consider in this framework derive from two simple IV regressions and a saturated IV regression. Specifically, we study the causal interpretation of the standard Wald estimand associated with treatment/instrument A conditional on $Z_B = 0$ and $Z_B = 1$, respectively, as well as the IV (2SLS) regression of the outcome on D_A, D_B and $D_A D_B$, instrumented by Z_A, Z_B and $Z_A Z_B$. There is now a substantial econometrics literature on causal inference in similar settings. Our results contribute to this growing body of work in the following ways.[1]

First, we employ weaker restrictions on instrument spillovers than other studies in the literature. Specifically, we allow for a compliance type for whom the presence of, say, Z_B represents a strong incentive against taking treatment A; we call this type the *cross-defiers*. At the same time, our framework also accommodates *joint compliers*—a type that reacts positively to the presence of the partner instrument and ultimately takes the given treatment if (and only if) both instruments are present. While identification results with joint compliers are available in other studies (e.g., Vazquez-Bare 2022), cross-defiers are typically ruled out, despite our application in Section 5 showing that this is also an empirically relevant type.

Second, our general identification results make explicit the difficulties that instrument spillovers cause in identifying "standalone" average treatment effects and, even more starkly, in separating treatment interaction from treatment effect heterogeneity. In particular, we provide the causal interpretation of the interaction term in the saturated IV regression, and show that a generally interesting local average interaction effect is confounded by terms that depend on how different the average effects of the two treatments are across various subgroups. These confounding terms however vanish if there are no interactive types or there is no treatment effect heterogeneity.

Third, given the general lack of interpretability of the interaction term in the IV regression, we also provide partial identification results—that is, bounds—for a parameter measuring the average interaction effect of the two treatments in a specific subgroup. We bound the interaction effect directly, based on its definition, as well as indirectly by bounding the confounding heterogeneity terms that show up in the interpretation of the interaction term. We further consider a formal Manski-type bounding strategy, which only uses the data and weak auxiliary assumptions, and a less formal strategy that relies on heuristic restrictions on treatment effect heterogeneity. The application illustrates all bounding approaches.

Fourth, while the results discussed in the main text impose one-sided noncompliance on the individual instruments (similarly to many results in the literature), we explore the consequences of relaxing this powerful but restrictive assumption in an appendix to the paper.[2] Specifically, we extend the analysis of the two Wald estimands to the case in which only one of the two instruments satisfies one-sided noncompliance, and even provide causal interpretations that do not require *any* monotonicity conditions.

The rest of the paper is organized as follows. In Section 2 we position our paper in the literature. Section 3 presents a formal potential outcome framework for pairs with endogenous (non-)compliance. We state and discuss our identification results in Section 4. Section 5 applies our theory in the context of Example 1. Section 6 concludes. There is a substantial amount

of supplementary material relegated to appendices. Appendix S1 provides supporting identification results. Appendix S2 explores the relaxation of one-sided non-compliance. Appendix S3 contains the proofs of all results in the main text, and finally, Appendix S4 supplements the empirical application.

2 | Related Literature

The seminal paper by Imbens and Angrist (1994) investigates endogenous non-compliance with a binary instrument, representing intended assignment to a binary treatment, when individual treatment effects are heterogeneous. The study establishes the now well-known result that, given weak monotonicity of the treatment in the instrument, a simple IV regression identifies the local average treatment effect (LATE) among compliers.

Several studies extend this framework to multiple treatments or instruments, which typically entails refinements of the monotonicity assumption. For example, Mogstad, Torgovitsky, and Walters (2020) impose partial monotonicity of the treatment in one instrument conditional on the other instrument(s), while Goff (2022) invokes vector monotonicity, which implies that each instrument affects treatment uptake in a direction that is common across subjects. Closer to our framework, Behaghel, Crépon, and Gurgand (2013) consider two mutually exclusive binary treatments along with two binary instruments. Ruling out instrument spillovers, they identify the LATEs among the two groups of compliers. Kirkeboen, Leuven, and Mogstad (2016) exploit information on next-best treatment alternatives to ease the assumption of no instrument spillovers. Heckman and Pinto (2018) allow for instrument spillovers but impose an unordered monotonicity assumption. Lee and Salanié (2018) discuss LATE identification without an unordered monotonicity assumption in the presence of sufficiently many continuous (rather than binary) instruments when treatment choice is governed by threshold-crossing models.

Another strand of related literature is concerned with relaxations of SUTVA, allowing for specific forms of interference among treatments, in various (quasi-)experimental settings; see, for instance, Sobel (2006), Hong and Raudenbush (2006), Hudgens and Halloran (2008), Ferracci, Jolivet, and van den Berg (2014), or Huber and Steinmayr (2021). Particularly relevant in our context are the studies by Kang and Imbens (2016), Imai, Jiang, and Malani (2021), and Vazquez-Bare (2022), who combine relaxations of SUTVA with treatment non-compliance. In addition, Blackwell (2017) studies the interaction between two randomized treatments with non-compliance with an application in political science. We now discuss the relationship between the last four papers and ours in more detail.

Kang and Imbens (2016) and Imai, Jiang, and Malani (2021) consider a partial interference framework, where interference between a given unit's outcome and a peer's treatment occurs within well-defined clusters, such as geographic regions. (In our setting the clusters are the pairs, i.e., there is only one peer.) In addition to conventional IV assumptions, Kang and Imbens (2016) impose a treatment exclusion restriction, ruling out coordinated compliance. This restriction permits identification of the direct LATE, that is, the average effect of the own treatment among compliers. Under an additional one-sided non-compliance assumption, one can also identify the average interference (spillover) effect of the peers' treatments, in the absence of the own treatment, within the whole population. Imai, Jiang, and Malani (2021) also study the identification of the direct LATE and present a condition that holds under the treatment exclusion restriction but is even satisfied under the weaker condition that a unit's treatment status does not depend on the instrument values of those peers who are non-compliers with respect to their own instruments. Our paper allows for more general forms of instrument spillovers and makes explicit the resulting difficulties in identifying easily interpretable causal effects.

The framework of our paper is most closely related to Blackwell (2017; henceforth, BW) and Vazquez-Bare (2022; henceforth, VB), but it extends each in certain directions.[3] BW does not explicitly consider (partial) interference, but a scenario where a unit's outcome may be affected by two interacting binary treatments, D_A and D_B, each with a distinct binary instrument, Z_A and Z_B, respectively. As discussed above, this setup can be viewed as a special case of our framework where pair members A and B are the same unit. Just as Kang and Imbens (2016), BW imposes a treatment exclusion restriction to identify the direct LATE of a given treatment, conditional on the other treatment being active or not, as well as the local average interaction effect of the two treatments, among complier units who follow the respective instruments in both their takeup decisions. Again, our framework and VB's are more general in that instrument spillovers are allowed; on the other hand, BW's results do not impose one-sided noncompliance.

VB explicitly considers a paired design and imposes a weaker first stage condition than the treatment exclusion restriction. Specifically, VB assumes that a pair member's potential treatment status indicators (say, $D_A(z_A, z_B)$) obey a specific ordering across the four possible instrument configurations; namely, $D_A(1,1) \geq D_A(1,0) \geq D_A(0,1) \geq D_A(0,0)$. In deriving his main results, VB also imposes one-sided non-compliance, which implies the last two inequalities above, and facilitates the identification of LATE among compliers with untreated peers as well as the spillover effect on untreated units induced by complying peers.

Importantly, our general results are derived under even weaker monotonicity restrictions on the instruments than in VB. While we also maintain one-sided non-compliance with respect to a treatment's "own" instrument, we do not impose monotonicity restrictions on how the partner instrument affects treatment takeup. As noted in the introduction, we accommodate a non-standard compliance type called cross-defiers, necessary for the application in Example 1, but not present in any other compliance framework we are aware of. In a supplement to this paper, we also investigate relaxations of one-sided noncompliance.

Given our general assumptions, we show that the Wald estimand for treatment A, conditional on $Z_B = 0$, reflects the average treatment effect in the union of two groups: compliers with Z_A and cross-defiers for Z_B. The second Wald estimand, conditional on $Z_B = 1$, has a complicated interpretation in general. However, under auxiliary conditions this estimand also lends itself to an insightful causal interpretation, given by the weighted average of three local average treatment effects. Similarly to our paper, both

BW and VB consider the saturated IV regression but under different sets of conditions. BW rules out instrument spillovers and assumes statistical independence between the instruments while VB does not allow for cross-defiers and also uses one-sided noncompliance. In BW's framework the coefficient on $D_A D_B$ does identify a meaningful local average interaction effect between the two treatments, while VB does not attempt to draw out any interesting causal parameters buried in this estimand.[4] It is true that in our general framework, this coefficient also does not lend itself to a "clean" interpretation. Instead, the interaction effect, often of central interest in applications, is bound up with terms that result from instrument spillovers and treatment effect heterogeneity across various compliance types. We state partial identification results that provide bounds for the interaction effect.

A final paper we must acknowledge here is a contribution by Bhuller and Sigstad (2024), who also consider a multiple treatment framework. However, the study differs from ours, as well as from BW and VB, in that it does not focus on identifying a specific LATE or interaction effect within a well-defined group of compliers. Instead, it aims to provide conditions under which a 2SLS regression with multiple treatments consistently estimates the weighted average effect of a given treatment across multiple compliance types, ensuring proper (in particular, non-negative) weights under arbitrary treatment effect heterogeneity. The authors demonstrate that it is necessary and sufficient to have an "average conditional monotonicity" and "no cross effects" condition hold, encompassing the treatment exclusion restriction (along with treatment monotonicity in the own instrument) as a special case. In Section 4, we will use more specific insights from this paper in explaining the structure of our own results.

3 | A Potential Outcome Framework for Pairs

3.1 | Variable Definitions

The population consists of ordered pairs of individuals (e.g., married couples); we will refer to the first member of a pair as member A and the second as member B. There are two potentially different binary treatments: D_A is targeted at member A and D_B is targeted at member B. By representing individual units as pairs with identical members, the setup also accommodates the analysis of two (interacting) treatments received by a single unit.

We are interested in the effect of D_A and/or D_B on some dependent variable Y. This outcome may be associated with member A alone, member B alone, or the pair itself. The observed value of Y is given by one of four potential outcomes: $Y(d_A, d_B)$ for $d_A, d_B \in \{0,1\}$. For example, $Y(1,0)$ is the potential outcome if one imposes $D_A = 1$ and $D_B = 0$; that is, member A is exposed to treatment A, but member B is not exposed to treatment B. To make the notation less cluttered, we will omit the comma and simply write $Y(10)$ whenever actual figures ("1" and/or "0") are used in the argument. Using the potential outcomes and the treatment status indicators, we can formally express the observed outcome as

$$\begin{aligned} Y = {} & Y(11)D_A D_B + Y(10)D_A(1-D_B) + Y(01)(1-D_A)D_B \\ & + Y(00)(1-D_A)(1-D_B). \end{aligned} \tag{1}$$

Treatment effect identification is facilitated by a pair of binary instruments, Z_A and Z_B, assigned to pair members A and B, respectively. We think of these instruments as indicators of (randomly assigned) treatment eligibility or the presence of an exogenous incentive to take the corresponding treatment. The leading example is a randomized control trial, where Z_A and Z_B are the experimenter's intended treatment assignments for pair member A and B, respectively. Compliance with these assignments is, however, endogenous and possibly coordinated across pair members. We refer to Z_A as member/treatment A's *own instrument* and Z_B as the *partner instrument*. The labels are of coursed reversed for treatment B.

Thus, there are four potential treatment status indicators associated with each pair member; they are denoted as $D_A(z_A, z_B)$ for member A and $D_B(z_A, z_B)$ for member B, $z_A, z_B \in \{0,1\}$. For example, $D_A(01)$ indicates whether member A of a pair takes up treatment A when they are not assigned ($Z_A = 0$) but their partner is assigned to treatment B ($Z_B = 1$). The actual treatment status of member A can be written as

$$\begin{aligned} D_A = {} & D_A(11)Z_A Z_B + D_A(10)Z_A(1-Z_B) + D_A(01)(1-Z_A)Z_B \\ & + D_A(00)(1-Z_A)(1-Z_B). \end{aligned} \tag{2}$$

There is of course a corresponding formula for D_B.

We now formally impose standard IV assumptions on Z_A and Z_B.

Assumption 1. [IV] (i) Given the values of the treatment status indicators D_A and D_B, the potential outcomes do not depend on the instruments Z_A and Z_B. (ii) The instruments (Z_A, Z_B) are jointly independent of the potential outcomes and the potential treatment status indicators. (iii) $P(Z_A = 1) \in (0,1)$, $P(Z_B = 1) \in (0,1)$ and $P(Z_A = Z_B) \in (0,1)$.

The exclusion restriction stated in part (i) of Assumption 1 is one of the defining properties of an instrument, and it justifies (ex-post) the potential outcomes being indexed by (d_A, d_B) only. Part (ii), known as "random assignment," states that the instrument values (Z_A, Z_B) are exogenously determined. This assumption holds, by design, in an experimental setting where intended treatment assignments are explicitly randomized. Part (iii) states that the intended treatment assignments follow a 2×2 factorial design; that is, there is a positive fraction of pairs assigned to each of the following four categories: treatment A alone, treatment B alone, both treatments, or neither treatment.

We will impose further assumptions on the potential treatment status indicators in Section 3.3.

3.2 | Parameters of Interest

Let $\mathcal{P}$ be a subset of the population of pairs. We define the following treatment effect parameters and notation:

- $ATE_{A|\overline{B}}(\mathcal{P}) := E[Y(10) - Y(00)|\mathcal{P}]$ denotes the average effect of applying treatment A alone in the subpopulation $\mathcal{P}$. In other words, this is the average effect of treatment A conditional on treatment B being "turned off" in the subpopulation $\mathcal{P}$.

- $ATE_{A|B}(\mathcal{P}) := E[Y(11) - Y(01)|\mathcal{P}]$ denotes the average effect of applying both treatments to the subpopulation $\mathcal{P}$ relative to applying treatment B alone; in other words, this is the average effect of treatment A conditional on maintaining treatment B.
- $ATE_{AB}(\mathcal{P}) := E[Y(11) - Y(00)|\mathcal{P}]$ denotes the average effect of applying treatment A and B jointly to the subpopulation $\mathcal{P}$ relative to applying no treatment at all.

The parameters $ATE_{A|\overline{B}}(\mathcal{P})$ and $ATE_{A|B}(\mathcal{P})$ are called local average conditional effects, or LACEs, by BW, while $ATE_{AB}(\mathcal{P})$ is called the local average joint effect (LAJE). For a given group $\mathcal{P}$, the difference between the two conditional effects measures the interaction between the two treatments within $\mathcal{P}$, and is hence termed the local average interaction effect (LAIE) by *ibid.* That is,

$$LAIE(\mathcal{P}) = ATE_{A|B}(\mathcal{P}) - ATE_{A|\overline{B}}(\mathcal{P}).$$

If the LAIE is positive, the two treatments reinforce each other, while if it is negative, then they work against each other. One can define analogous LACE parameters for treatment B by interchanging the roles of A and B in the definitions above. The associated joint and interaction effects stay unchanged.

In case the pairs have distinct members, the interpretation of these parameters also depends on the definition of the outcome Y. In particular, if Y is associated with pair member A alone, then $ATE_{A|\overline{B}}(\mathcal{P})$ and $ATE_{A|B}(\mathcal{P})$ measure what is called the direct effect of treatment A by Hudgens and Halloran (2008). On the other hand, if Y is associated with member B alone, then $ATE_{A|\overline{B}}(\mathcal{P})$ and $ATE_{A|B}(\mathcal{P})$ measure the indirect or spillover effect of treatment A on pair member B. For example, if the treatment is vaccination, and the outcome is the incidence of a disease, then the vaccination of member A confers protection on member A, but also indirectly protects his or her partner.

3.3 | Compliance Types

The setup presented in Section 3.1 assigns four potential treatment indicators to each pair member, corresponding to the four possible incentive schemes represented by (Z_A, Z_B). Without any further restrictions on treatment takeup, the possible configurations of these 8 potential treatment variables partition the population of pairs into $2^8 = 256$ different compliance profiles. At this level of generality a couple of regression-based estimands can hardly be a meaningful summary of the various average treatment effects across types. Therefore, similarly to VB, we impose one-sided noncompliance with respect to the treatment's own instrument, which dramatically reduces the number of possible compliance profiles.

Assumption 2. [One-sided noncompliance] (i) $D_A(0, z) = 0$ and (ii) $D_B(z, 0) = 0$ for $z \in \{0,1\}$.

Assumption 2 states that neither member of the pair has access to their own treatment unless they have been "randomized in"; that is, the value of their own instrument is 1. In other words, one-sided noncompliance presumes that the experimenter is able to exclude individuals from all sources of the treatment. Whether or not this assumption is reasonable depends on the institutional setting and details of the underlying experiment, but it often fails in practice. Therefore, we consider relaxations of Assumption 2 in Appendix S2.

Under Assumption 2, each pair member may belong to one of only four compliance types, summarized by the following definition.

Definition 1. Under Assumption 2, member A of a pair (A, B) is one of four compliance types:

	$D_A(00)$	$D_A(01)$	$D_A(10)$	$D_A(11)$
Self-complier (s)	0	0	1	1
Joint complier (j)	0	0	0	1
Never taker (n)	0	0	0	0
Cross-defier (d)	0	0	1	0

Furthermore, a pair member is a *complier (c)* if they are either a self-complier or joint-complier.

Remarks

1. The corresponding definitions for member B can be obtained by interchanging the two arguments of the potential treatment status indicators, while using the subscript B.
2. A self-complier's treatment status is determined solely by the value of their own instrument. By contrast, a joint complier takes the treatment if and only if both instruments are turned on; their own instrument is not sufficient to induce participation. (VB's terminology is group compliers.)
3. Cross-defiers are a non-standard type. If member A is a cross-defier, then they will comply with their own instrument Z_A as long as the other instrument is absent. However, for such individuals the presence of Z_B represents a strong incentive *against* takeup of D_A; so strong in fact that it overpowers the presence of Z_A and causes the individual to abandon treatment. Thus, the individual A acts in defiance of $Z_B = 1$.[5]
4. Finally, a never taker cannot be induced to take the treatment by any instrument configuration.

Given the four individual compliance types, every pair belongs to one of the 16 compliance profiles $\{s, j, d, n\} \times \{s, j, d, n\}$. For example, (s, j) is the set of pairs where A is a self-complier and B is a joint complier, and so forth. Furthermore, we will use the notation $(c, \cdot)$ to denote the set of pairs where member A is a complier, and so forth, and $P(s, n)$ to denote the probability that for a randomly drawn pair, A is a self-complier and B is a never-taker, and so forth.

The following assumption ensures that some of the compliance categories are not vacuous (e.g., there are at least some individuals who respond to their own instrument).

Assumption 3. [First stage] (i) $P(D_A(10) = 1) > 0$ and $P(D_B(01) = 1) > 0$; (ii) $P(D_A(11) = 1) > 0$ and $P(D_B(11) = 1) > 0$.

Part (i) of Assumption 3 means that $P(s \cup d, \cdot) > 0$ and $P(\cdot, s \cup d) > 0$ while part (ii) means $P(c, c) > 0$. These conditions ensure that the IV estimands considered in Section 4 are well defined.

A testable implication of the existence of joint compliers (with respect to treatment A) is that if one runs a simple OLS regression of D_A on Z_B in the $Z_A = 1$ subsample, then the coefficient of Z_B should be positive. Conversely, if the coefficient is negative, then cross-defiers must be present in the population.[6] Clearly, joint compliers treat the presence of Z_B as a positive incentive to take treatment A, while cross-defiers treat it as a (strong) disincentive. Given the situation, it may be possible to argue that the two behaviors do not exist simultaneously; that is, one could rule out joint compliers or cross-defiers for treatment A, depending on which type is not needed to explain the sign of the regression coefficient. Our empirical application in Section 5 illustrates how to exploit such simplifications in practice to enhance the interpretation of IV estimands.

Our framework can also accommodate simplifying assumptions on pair formation. For example, one might postulate that there are no (n, j) or (j, n) pairs. This assumption is plausible if member A's utility of taking treatment A is affected by Z_B only through member B's actual treatment status D_B, which more formally means that $D_A(z_A, z_B)$ is of the form $f(z_A, D_B(z_A, z_B))$. If B is a never taker then $D_B = 0$, and the value of Z_B is not relevant for A's decision. Hence A cannot be a joint complier. For example, Z_B could be a randomized monetary reward payable only on actual takeup of treatment B. If B is never treated, the reward is not paid out and should be irrelevant to A. On the other hand, (n, j) or (j, n) pairs may well exist if A has direct access to the incentive represented by Z_B. This is the case, for example, if the pair stands for a single unit targeted by two treatments.[7]

4 | Identification Results

4.1 | Population Proportion of Compliance Profiles

The exact type of a given pair is generally unobserved as it depends on the pair's behavior in counterfactual scenarios. Nevertheless, the observed conditional probabilities

$$P(D_A = d_A, D_B = d_B | Z_A = z_A, Z_B = z_B),\ d_A, d_B, z_A, z_B \in \{0,1\} \tag{3}$$

can be used to identify the relative frequency of a number of compliance profiles in the population. Nevertheless, not all probabilities under (3) carry independent information. This is for two reasons: first, for any given (z_A, z_B), the corresponding probabilities add up to 1, and, second, $Z_A = 0$ automatically implies $D_A = 0$, and $Z_B = 0$ automatically implies $D_B = 0$ by Assumption 2. It follows that there are only five independently informative moments, which of course makes it impossible to identify the relative frequencies of all 16 compliance profiles separately. Lemma 1 presents the interpretation of five selected conditional probabilities that are linearly independent.

Lemma 1. Suppose that Assumptions 1 and 2 are satisfied. Then:

$$\begin{aligned}
P(D_A = 1 | Z_A = 1, Z_B = 0) &= P(s \cup d, \cdot) = P(s, \cdot) + P(d, \cdot) \\
P(D_B = 1 | Z_A = 0, Z_B = 1) &= P(\cdot, s \cup d) = P(\cdot, s) + P(\cdot, d) \\
P(D_A = 1, D_B = 0 | Z_A = 1, Z_B = 1) &= P(c, n \cup d) = P(c, n) + P(c, d) \\
P(D_A = 0, D_B = 1 | Z_A = 1, Z_B = 1) &= P(n \cup d, c) = P(n, c) + P(d, c) \\
P(D_A = 1, D_B = 1 | Z_A = 1, Z_B = 1) &= P(c, c).
\end{aligned}$$

In consequence, the compliance profile shares $P(n \cup d, n \cup d)$, $P(j \cup n, \cdot)$, $P(c, \cdot)$ and $P(n \cup d, \cdot)$ are also identified (along with the symmetric expressions obtained by interchanging the role of the pair members); see Corollary S1.1 in Appendix S1.

4.2 | The Causal Interpretation of Three IV Estimands

Each pair in the target population is associated with an observed 5-vector (Y, D_A, D_B, Z_A, Z_B). Given a sample of observations on this vector, one may run several different IV regressions using the full sample or a suitable subsample.

i. Consider the IV regression of Y on D_A and a constant in the $Z_B = 0$ subsample, using Z_A as an instrument for D_A. Under general conditions, the Wald estimand

$$\delta_{A0} = \frac{E(Y | Z_A = 1, Z_B = 0) - E(Y | Z_A = 0, Z_B = 0)}{E(D_A | Z_A = 1, Z_B = 0) - E(D_A | Z_A = 0, Z_B = 0)}. \tag{4}$$

represents the probability limit of the slope coefficient associated with D_A.

ii. The IV regression described in point (i) above can also be implemented in the $Z_B = 1$ subsample. Under general conditions, the Wald estimand

$$\delta_{A1} = \frac{E(Y | Z_A = 1, Z_B = 1) - E(Y | Z_A = 0, Z_B = 1)}{E(D_A | Z_A = 1, Z_B = 1) - E(D_A | Z_A = 0, Z_B = 1)}. \tag{5}$$

represents the probability limit of the slope coefficient associated with D_A.

iii. One can also run a full-sample IV regression of Y on a constant, D_A, D_B, and $D_A D_B$, instrumented by Z_A, Z_B and $Z_A Z_B$. More formally, let $D = (D_A, D_B, D_A D_B)'$, $\ddot{D} = (1, D')'$, $Z = (Z_A, Z_B, Z_A Z_B)'$ and $\ddot{Z} = (1, Z')'$. Under general conditions, the full-sample IV estimator is a 4×1 vector that converges to the estimand

$$\beta = (\beta_0, \beta_A, \beta_B, \beta_{AB})' = [E(\ddot{Z}\ddot{D}')]^{-1} E(\ddot{Z}Y).$$

The following three theorems state the causal interpretation of these estimands. The proofs are provided in Appendix S3.

Theorem 1. Under Assumptions 1, 2, and 3, the Wald estimand (4) satisfies

$$\begin{aligned}
\delta_{A0} = ATE_{A|\overline{B}}(s \cup d, \cdot) = ATE_{A|\overline{B}}(s, \cdot) \frac{P(s, \cdot)}{P(s \cup d, \cdot)} \\
+ ATE_{A|\overline{B}}(d, \cdot) \frac{P(d, \cdot)}{P(s \cup d, \cdot)}.
\end{aligned}$$

Remarks

1. Theorem 1 states that the Wald estimand δ_{A0} identifies the average effect of treatment A alone among pairs where member A is a self-complier or a cross-defier. If the latter type is not present, the estimand reduces to the "classic" LATE parameter.
2. In some applications Y could be an outcome associated solely with member B. In this case δ_{A0} identifies the average *spillover* effect on member B of a treatment applied to member A — among pairs where A is a self-complier or cross-defier.
3. Given that $Z_B = 0$, Z_A has no effect on D_B because one-sided noncompliance forces $D_B = 0$. Furthermore, Z_A has a (weakly) positive effect on D_A since member A cross-defiers also comply with Z_A when $Z_B = 0$. These facts in our framework correspond to the "no cross effects" and "average conditional monotonicity" requirements by Bhuller and Sigstad (2024). As *ibid*. show, it is precisely under these conditions that IV regression coefficients in a multiple treatment setting recover a properly weighted average treatment effect across compliance types, just as in Theorem 1.

In Appendix S2, we consider an extension of Theorem 1 to the case in which Z_B continues to satisfy one-sided noncompliance but Z_A only obeys a general monotonicity condition. The result is similar to Theorem 1 except that an additional compliance profile must be added to the set of pairs $(s \cup d, \cdot)$; namely, pairs where member A is a "cross-complier" in the sense that they take the treatment in response to any one of the instruments being turned on (possibly the partner instrument Z_B alone). The extension is in fact derived from a very general representation theorem that gives causal interpretations to δ_{A0} and δ_{A1} without imposing *any* monotonicity conditions on the instruments. This latter result is rather too general to be useful in practice; its main value lies in the fact that it can readily be "customized" via auxiliary restrictions that fit the application at hand. For example, the general result also shows that the extension of Theorem 1 to the case in which Z_B does not satisfy one-sided noncompliance is much more complicated, even when Z_A does obey this restriction.

The next result states the causal interpretation of δ_{A1}.

Theorem 2. Under Assumption 1, 2, and 3, the Wald estimand (5) satisfies

$$\begin{aligned}\delta_{A1} = {} & ATE_{AB}(c,j)\frac{P(c,j)}{P(c,\cdot)} + ATE_{A|B}(c,s)\frac{P(c,s)}{P(c,\cdot)} \\ & + ATE_{A|\overline{B}}(c,n\cup d)\frac{P(c,n\cup d)}{P(c,\cdot)} \\ & - ATE_{B|\bar{A}}(\cdot,d)\frac{P(\cdot,d)}{P(c,\cdot)} + ATE_{B|\bar{A}}(n\cup d,j)\frac{P(n\cup d,j)}{P(c,\cdot)}\end{aligned} \tag{6}$$

Corollary 1. Suppose, in addition, that there are no cross-defiers with respect to either instrument and there are no (n,j) pairs. Then the Wald estimand (5) is equal to

$$\begin{aligned}\delta_{A1} = {} & ATE_{AB}(c,j)\frac{P(c,j)}{P(c,\cdot)} + ATE_{A|B}(c,s)\frac{P(c,s)}{P(c,\cdot)} \\ & + ATE_{A|\overline{B}}(c,n)\frac{P(c,n)}{P(c,\cdot)}.\end{aligned} \tag{7}$$

Remarks

1. In Appendix S2, we present an extension of Theorem 2 to the case in which Z_A continues to satisfy one-sided noncompliance but Z_B only obeys a weaker monotonicity condition.
2. Again, the interpretation of Theorem 2 and Corollary 1 is enriched by the fact that Y can be an outcome associated with A alone, B alone, or the pair (A,B).
3. Given that $Z_B = 1$, there is a much richer set of possible responses to Z_A, making the interpretation of δ_{A1} complicated. For example, Z_A now affects the takeup of D_B, positively for member B joint compliers and negatively for cross-defiers. Again, the high-level results by Bhuller and Sigstad (2024) show that in this case IV estimands in a multiple treatment setting generally conflate the effects of the various treatments with partly negative weights, just as in Theorem 2.
4. The simplifying assumptions imposed in Corollary 1 are motivated and discussed in Section 3.3.

To understand the causal effects appearing in the special case (7), consider changing Z_A from 0 to 1 conditional on $Z_B = 1$. In this case, (c,j) pairs will switch from no treatment at all to both treatments, contributing the first term in (7). For (c,s) pairs, member A switches from no treatment to treatment A, while member B continues to take treatment B throughout. This contributes the second term. Finally, among (c,n) pairs, member A switches from no treatment to treatment A, while member B continues to abstain from treatment. This option contributes the last term. As in the absence of cross-defiers $P(c,\cdot) = P(c,j) + P(c,s) + P(c,n)$, the probability weights in (7) sum to one and are identified from the observed data using Lemma 1 and Corollary S1.1. The general expression (6) follows the same logic — it reflects the reaction of various types of pairs to changing Z_A from 0 to 1 while maintaining $Z_B = 1$. However, there are now more possibilities, including member B cross-defiers dropping D_B when Z_A is turned on. The end result is a linear combination of average treatment effects where some of the weights are negative and do not sum to one.

Finally, Theorem 3 states the causal interpretation of the elements in the coefficient vector $\beta = (\beta_0, \beta_A, \beta_B, \beta_{AB})'$.

Theorem 3. Given Assumptions 1, 2 and 3, $\beta_0 = E[Y(00)]$, $\beta_A = ATE_{A|\overline{B}}(s\cup d,\cdot)$ and $\beta_B = ATE_{B|\bar{A}}(\cdot, s\cup d)$. In addition,

$$\beta_{AB} = ATE_{A|B}(c,c) - ATE_{A|\overline{B}}(c,c) \tag{8}$$

$$+\frac{P(j,\cdot)}{P(c,c)}[ATE_{A|\overline{B}}(j,\cdot) - ATE_{A|\overline{B}}(s\cup d,\cdot)] \tag{9}$$

$$+\frac{P(\cdot,j)}{P(c,c)}[ATE_{B|\bar{A}}(\cdot,j) - ATE_{B|\bar{A}}(\cdot, s\cup d)] \tag{10}$$

$$+\frac{P(d,\cdot)}{P(c,c)}[ATE_{A|\overline{B}}(s\cup d,\cdot)-ATE_{A|\overline{B}}(d,\cdot)] \tag{11}$$

$$+\frac{P(\cdot,d)}{P(c,c)}[ATE_{B|\bar{A}}(\cdot,s\cup d)-ATE_{B|\bar{A}}(\cdot,d)]. \tag{12}$$

Remarks

1. The coefficients on the stand-alone treatment dummies are the same as the split-sample Wald estimands that condition on the partner instrument being zero.
2. The coefficient on the interaction term has a complex interpretation. Term (8) is the local average interaction effect (LAIE) of the two treatments among (c,c) pairs, which would presumably be of interest in many applications. However, this quantity is confounded by additional terms that depend on the heterogeneity of the average treatment effect across types. For example, term (9) compares the average effect of treatment A, applied in isolation, across two subpopulations: pairs where member A is a joint complier versus pairs where member A is a self-complier or a cross-defier. While the latter average treatment effect is identified by β_A, the former is not. Terms (10), (11) and (12) have analogous interpretations.
3. The presence of the confounding "heterogeneity terms" is due to joint compliers and cross-defiers—interactive types that react to their partner's instrument as well. Under the treatment exclusion restriction these types are not present, and β_{AB} identifies the LAIE of the two treatments among (s,s) pairs, as also shown by Theorem 2 of BW.[8]

4.3 | Partial Identification of LAIE

The coefficient β_{AB} in Theorem 3 does not have a clean interpretation because it conflates the interaction between the two treatments with the heterogeneity of the treatment effects across various compliance types. We now show that it is still possible to learn about $LAIE(c,c)$ through bounds constructed under some auxiliary conditions. There are two different approaches. First, it is possible to bound $LAIE(c,c)$ directly, based on the moments in its definition. Second, one can take the causal interpretation of β_{AB} in Theorem 3 as a starting point, and bound the influence of the heterogeneity terms (9) through (12). In doing so, one obtains an indirect bound on $LAIE(c,c)$ as well. We present the direct bounds here in the main text; the indirect ones are stated in Appendix S4. The application in Section 5 illustrates both approaches.

There are four conditional means involved in the definition of $LAIE(c,c)$:

$$\begin{aligned}&E[Y(11)|(c,c)],\ E[Y(00)|(c,c)],\\ &E[Y(10)|(c,c)]\text{and }\ E[Y(01)|(c,c)].\end{aligned} \tag{13}$$

The first quantity under (13) is identified directly from the data by the conditional expectation of Y given $D_A=D_B=1$ and $Z_A=Z_B=1$; see Lemma S1.2 in Appendix S1. We bound the remaining moments in the spirit of Manski (1989, 1990), using the following assumption.

Assumption 4. [bounds] $(i)\ Y(d_A,d_B)\in[0,K]$ for some $K>0$; $(ii)\ Y(10)\geq Y(00)$ and $Y(01)\geq Y(00)$.

Part (i) states that the potential outcomes are bounded; the fact that the lower bound is set to zero is a normalization. Part (ii) postulates that when treatment A is applied in isolation, it has a positive effect on any individual unit, and the same is assumed about treatment B. While researchers often hold prior expectations about the sign of an average treatment effect, the requirement that the sign applies uniformly in the population is a non-trivial homogeneity restriction known as monotone treatment response (Manski 1997). Importantly, however, part (ii) does not restrict the sign of the interaction effect.

As a first step toward bounding the interaction effect, we provide bounds for the joint effect of the two treatments.

Theorem 4. Suppose that Assumptions 1 through 4 are satisfied. Then the following inequalities hold true:

a. $L_{00}(c,c)\leq E[Y(00)|(c,c)]\leq U_{00}(c,c)$, where

$$\begin{aligned}U_{00}(c,c)=&E[Y(00)]\frac{1}{P(c,c)}\\ &-E[Y(00)|(n\cup d,n\cup d)]\frac{P(n\cup d,n\cup d)}{P(c,c)},\\ L_{00}(c,c)=&U_{00}(c,c)-E[Y(10)|(c,n\cup d)]\frac{P(c,n\cup d)}{P(c,c)}\\ &-E[Y(01)|(n\cup d,c)]\frac{P(n\cup d,c)}{P(c,c)},\end{aligned}$$

and each probability and expectation in the definition of $L_{00}(c,c)$ and $U_{00}(c,c)$ is identified from the data as specified by Lemma 1 and Lemma S1.2.

b. In consequence,

$$\begin{aligned}&E[Y(11)|(c,c)]-U_{00}(c,c)\leq ATE_{AB}(c,c)\\ &\leq E[Y(11)|(c,c)]-L_{00}(c,c),\end{aligned} \tag{14}$$

where $E[Y(11)|(c,c)]$ is identified as in Lemma S1.2.

Remarks

1. Lemma S1.2 in Appendix S1 provides the causal interpretation of all conditional moments $E[Y|D_A=d_A,D_B=d_B,Z_A=z_A,Z_B=z_B]$, $d_A,d_B,z_A,z_B\in\{0,1\}$.
2. The proof of Theorem 4 is given in Appendix S3; the construction is similar to Manski's classic work cited above. We expand $E[Y(00)]$ as a weighted average of $E[Y(00)|(c,c)]$ and three other conditional expectations over different subgroups. One of the latter expectations is point-identified from the data, and, under Assumption 4, the other two expectations can be bounded by identified ones (from above) and zero (from below). We rearrange the resulting inequalities to obtain bounds for $E[Y(00)|(c,c)]$.

3. If $L_{00}(c,c)$ is negative, it may be replaced by zero; if $U_{00}(c,c)$ is greater than K, it may be replaced by K. If no such replacements are made, then one can in principle apply the classic theory in Imbens and Manski (2004) to construct confidence intervals for the partially identified parameter $E[Y(00)|(c,c)]$. In this paper we focus on identification and forego further discussion of inference.

4. If one strengthens Assumption 4 to include the condition $Y(11) \geq Y(00)$, then an improved upper bound to $E[Y(00)|(c,c)]$ is given by the minimum of $U_{00}(c,c)$ and $E[Y(11)|(c,c)]$, and the joint treatment effect cannot be less than zero. This extra assumption is not entirely trivial but can still be plausible in applications (see Section 5.2).

The relationship between the average joint and interaction effect can be written as

$$\begin{aligned} LAIE(c,c) = ATE_{AB}(c,c) - ATE_{A|\overline{B}}(c,c) \\ - ATE_{B|\overline{A}}(c,c). \end{aligned} \tag{15}$$

While the average effects of treatments A and B, applied in isolation, are not identified in the (c,c) subgroup, they are identified in subgroups $(s \cup d, \cdot)$ and $(\cdot, s \cup d)$, respectively (see Theorem 3). It is natural to take the identified subgroup effects as a reference point and speculate about other groups on the basis of these. In particular, we may write $ATE_{A|\overline{B}}(c,c) = \lambda_A \cdot ATE_{A|\overline{B}}(s \cup d, \cdot)$ for some (unknown) multiplier $\lambda_A \geq 0$, and define λ_B similarly for treatment B. Combining these expressions with (14) and (15) bounds the interaction effect in terms of λ_A and λ_B. One can then consider various hypotheses about these parameters; for example, it may be reasonable to postulate in a given application that $ATE_{A|\overline{B}}(c,c)$ is at most three times as large as $ATE_{A|\overline{B}}(s \cup d, \cdot)$, implying $\lambda_A \in [0,3]$. Or one may plot those (λ_A, λ_B) pairs for which the upper bound of LAIE is zero, etc. Boundaries of this type are common in the econometrics literature on sensitivity analysis (e.g., Masten and Poirier 2020; Martínez-Iriarte 2021). We demonstrate the construction and use of such heuristic bounds in the context of our application in Sections 5.2 and 5.3.

One can bound the interaction effect in a more formal way by also bounding the last two conditional means under (13). For these bounds to be potentially tighter than the interval $[0, K]$, we impose further compliance type restrictions. Motivated by the discussion following Assumption 2 in Section 3.3, we assume that treatment A does not admit cross defiers while treatment B does not admit joint compliers. (We will argue that such a restrictions are reasonable in our application, at least as a polar case.) This leads to the following result.

Theorem 5. Suppose that Assumptions 1 through 4 are satisfied. If, in addition, there are no $(d,\cdot)$ pairs and no $(\cdot,j)$ pairs, then the following inequalities hold true:

a. $L_{10}(c,c) \leq E[Y(10)|(c,c)] \leq U_{10}(c,c)$, where

$$\begin{aligned} L_{10}(c,c) =& E[Y(10)|(s,\cdot)]\frac{P(s,\cdot)}{P(c,c)} \\ & - E[Y(10)|(c,n \cup d)]\frac{P(c,n \cup d)}{P(c,c)} \\ U_{10}(c,c) =& L_{10}(c,c) + K\frac{P(j,\cdot)}{P(c,c)}, \end{aligned}$$

b. $L_{01}(c,c) \leq E[Y(01)|(c,c)] \leq U_{01}(c,c)$, where

$$\begin{aligned} U_{01}(c,c) =& E[Y(01)|(\cdot,s \cup d)]\frac{P(\cdot,s \cup d)}{P(c,c)} \\ & - E[Y(01)|(n,c)]\frac{P(n,c)}{P(c,c)}, \\ L_{01}(c,c) =& U_{01}(c,c) - K\frac{P(\cdot,d)}{P(c,c)}, \end{aligned}$$

and each probability and expectation in the definition of $L_{10}(c,c)$, $U_{10}(c,c)$, $L_{01}(c,c)$ and $U_{01}(c,c)$ is identified from the data as specified by Lemma 1 and Lemma S1.2.

c. In consequence,

$$\begin{aligned} E[Y(11)|(c,c)] + L_{00}(c,c) - U_{10}(c,c) - U_{01}(c,c) \\ \leq LAIE(c,c) \leq E[Y(11)|(c,c)] \\ + U_{00}(c,c) - L_{10}(c,c) - L_{01}(c,c), \end{aligned}$$

where $E[Y(11)|(c,c)]$ is identified as in Lemma S1.2.

5 | Empirical Application

5.1 | Data, Compliance Patterns, and IV Estimates

In this section, we present an empirical illustration of our theory based on data from the Student Achievement and Retention Project first analyzed by Angrist, Lang, and Oreopoulos (2009). This program, implemented on a campus in Canada in Fall 2005, randomly assigned two treatments, namely, academic services (tutoring) and financial incentives among first year college students whose high school grade point average was lower than the upper quartile. Tutoring included both access to more experienced students trained to provide academic support, as well as sessions aiming at improving study habits. The financial incentives consisted of conditional cash payments, ranging from 1000 to 5000 Canadian dollars, which were paid out if a student reached a specific average grade target. In our application, D_A is a binary variable indicating the takeup of any form of tutoring, while D_B is an indicator for signing up to receive financial incentives. We are interested in the impact of these treatments on the average grade at the end of the fall semester, which is our outcome variable Y, measured on a 0–100 scale.

The random offer of the treatments in the project was partly overlapping in the sense that some students were invited to either one of the treatments, to both, or neither. The instruments Z_A and Z_B correspond to binary indicators for being invited (and thus, being eligible) for tutoring and financial incentives. Thus, we are in the

special case of our framework where the very same individual is targeted by up to two distinct treatments, rather than having a pair of individuals that might be targeted by the same or separate treatments.

Treatment takeup D_A and D_B may endogenously differ from the random assignment Z_A and Z_B, respectively, because unobserved background characteristics such as personality traits likely drive both the treatment decision and academic performance. For instance, among those students offered tutoring and/or financial incentives, less motivated individuals satisfied with lower exam grades might not be willing to take the treatment(s), regardless of having received an offer or not. Due, in part, to such never takers, not all subjects comply with the random assignment, and the groups taking and not taking the treatment(s) generally differ in terms of outcome-relevant characteristics. On the other hand, among those students not offered the respective treatment, nobody managed to circumvent the assignment and take that treatment anyway. For this reason, non-compliance in our data is one-sided, as postulated in Assumption 2.

Applying traditional IV approaches (ruling out relaxations of the treatment exclusion restriction), the findings of Angrist, Lang, and Oreopoulos (2009) point to positive effects of the financial incentive or the combined treatments among females, but not among males. For this reason, our empirical illustration here only focuses on female students, leaving all in all 948 observations. However, for 150 females the outcome is missing, implying that these students did not take any exams in the fall semester. As the missing outcomes indicator is not statistically significantly associated with Z_A or Z_B (with p-values exceeding 20%), we drop these students from the sample, leaving us with 798 observations.

Table 1 provides descriptive statistics for our evaluation sample, namely the treatment and outcome means in the total sample and in the subsamples defined by the instrument values Z_A and Z_B. We see that the treatment frequencies observed in the data are consistent with one-sided noncompliance. As a further observation, the average grade (Y) is highest among female students receiving both instruments and lowest among those receiving neither. The difference in average outcomes between the two groups is statistically significant at the 1% level, pointing to a non-zero reduced form effect of the joint instruments Z_A and Z_B on Y. Moreover, the average outcome is somewhat higher among students exclusively eligible for financial incentives than among those exclusively eligible for tutoring (but this difference is not statistically significant at the 10% level).

To present the effect of the instruments on the treatments (the "first stage"), Tables 2, 3, and S4.1 report conditional probabilities of the first, second, and joint treatments, respectively, and relate them to specific compliance types. In analyzing treatment takeup patterns, we impose the restriction that there are no cross-defier types in the tutor treatment arm. We consider a similar restriction—no joint compliers—with respect to the financial incentive treatment, but we are more agnostic about this condition and impose it selectively in our bounding exercise later on.

Table 2 shows the takeup statistics for D_A. The absence of cross-defiers means that when being eligible for it, nobody is discouraged from actually taking up tutoring services by additionally being offered financial incentives. We see from Table 2 that the nonexistence of cross-defiers is consistent with the data since our estimates suggest that $P(D_A = 1|Z_A = 1, Z_B = 1) - P(D_A = 1|Z_A = 1, Z_B = 0) > 0$ (statistically significant at the 1% level). Ruling out cross-defiers is plausible if one agrees that, if anything, financial incentives for good grades should encourage (rather than discourage) the takeup of tutoring given that the latter is expected to increase academic performance. Without cross-defiers, the estimated shares of self-compliers (taking tutoring if and only if eligible for it), joint compliers (taking tutoring if and only if eligible for both treatments) and never takers amount to 28%, 21%, and 51%, respectively.

Similarly, Table 3 shows the takeup statistics for D_B. Not surprisingly, 93% sign up for the conditional payment when eligible for it (and nothing else). However, the estimates also suggest that $P(D_B = 1|Z_A = 1, Z_B = 1) < P(D_B = 1|Z_A = 0, Z_B = 1)$, the 12pp difference being statistically significant at the 5% level. For this inequality to hold, cross-defiers must be present in the population, and the prevalence of joint compliers must be limited. Cross-defiers behave oddly in that they accept the financial incentive if this is the only treatment they are eligible for, but they refuse it if they are additionally eligible for tutoring. While it is not clear why the availability of tutoring should be a disincentive for taking the conditional payment, these types are clearly present in the data.

Some of the subsequent analysis is simpler and more informative under the restriction that there are no joint compliers in the financial incentive treatment arm. Table 3 shows that the *combined* share of never-takers and joint compliers is estimated to be only 7%. Still, ruling out joint compliers altogether is a rather strong assumption, since this implies that access to tutoring cannot positively affect sign-up decisions for the financial incentive

TABLE 1 | Treatment and outcome means in the sample and by instruments.

Variable	Total sample	$Z_A = 1$ $Z_B = 1$	$Z_A = 0$ $Z_B = 1$	$Z_A = 1$ $Z_B = 0$	$Z_A = 0$ $Z_B = 0$
D_A (tutor)	0.08	0.49	0.00	0.28	0.00
D_B (fin. incentive)	0.22	0.81	0.93	0.00	0.00
Y (GPA)	63.78	66.98	65.75	63.57	62.83
Number of obs.	798	67	134	116	481

Note: Data from the Student Achievement and Retention Project; see Angrist, Lang, and Oreopoulos (2009). Female students only; those with missing GPA are dropped. GPA is measured on a 0–100 scale.

TABLE 2 | Conditional probabilities of treatment D_A.

Conditional probability	Estimate	Interpretation if no $(d, \cdot)$
$P(D_A = 1 \mid Z_A = 1, Z_B = 0)$	0.28	$P(s, \cdot)$
$P(D_A = 1 \mid Z_A = 1, Z_B = 1)$	0.49	$P(c, \cdot) = P(j, \cdot) + P(s, \cdot)$
$P(D_A = 1 \mid Z_A = 1, Z_B = 1)$		
$-P(D_A = 1 \mid Z_A = 1, Z_B = 0)$	0.21	$P(j, \cdot)$
$P(D_A = 0 \mid Z_A = 1, Z_B = 1)$	0.51	$P(n, \cdot)$

Note: D_A is the tutor treatment.

TABLE 3 | Conditional probabilities of treatment D_B.

Conditional probability	Estimate	Interpretation
$P(D_B = 1 \mid Z_A = 0, Z_B = 1)$	0.93	$P(\cdot, s \cup d) = P(\cdot, s) + P(\cdot, d)$
$P(D_B = 1 \mid Z_A = 1, Z_B = 1)$	0.81	$P(\cdot, c) = P(\cdot, s) + P(\cdot, j)$
$P(D_B = 1 \mid Z_A = 0, Z_B = 1)$		
$-P(D_B = 1 \mid Z_A = 1, Z_B = 1)$	0.12	$P(\cdot, d) - P(\cdot, j)$
$P(D_B = 0 \mid Z_A = 0, Z_B = 1)$	0.07	$P(\cdot, n) + P(\cdot, j)$

Note: D_B is the financial incentive treatment.

TABLE 4 | IV regression of Y on D_A, D_B, and $D_A D_B$.

Variable	Coefficient estimate	Standard error	p-value
Constant	62.83	0.55	0.00
D_A (tutor)	2.58	4.35	0.55
D_B (fin. incentive)	3.15	1.24	0.01
$D_A D_B$	0.69	5.31	0.90

Note: Y is GPA on a 0–100 grading scale. The instruments are the randomized treatment eligibility dummies and their interaction.

given eligibility for the latter. This would be violated if some individuals judged their chances of obtaining good grades to be highly dependent on tutoring, so much so that they would not even sign up for the financial incentive without access to tutoring. This behavior actually seems more reasonable than that of the cross-defiers. On the other hand, $P(\cdot, d) - P(\cdot, j) = 0.12$, so the higher the share of joint compliers, the higher the share of cross-defiers must be to explain the data. Given the odd behavior of the latter type, one could argue that their assumed prevalence should be as low as possible, which happens when there are no joint compliers.

There are additional moments of the data, which we present in Appendix S4. Specifically, Table S4.1 shows the joint distribution of D_A and D_B, conditional on eligibility for both treatments. These probabilities identify the shares of specific joint compliance profiles in the population, in accordance with Lemma 1 and Corollary S1.1. For example, 49% of the female students are estimated to have a (c, c) profile, meaning that they either comply with the intended assignment in both treatment arms, or they take both treatments when, and only when, they are eligible for both. In addition, Table S4.2 shows the average GPA conditional on all configurations of the treatment dummies and their instruments. As shown by Lemma S1.2 in Appendix S1, these moments identify the mean outcome for various compliance profiles.

Finally, in Table 4, we provide the results from the saturated two stage least squares regression studied in Theorem 3. The constant term provides an estimate for the mean potential outcome $E[Y(00)]$, suggesting that the average grade amounts to 62.83 points when female students neither take up tutoring, nor sign up for financial incentives. In the absence of cross-defiers, the estimate of β_A corresponds to the average effect of tutoring among self-compliers when the other treatment is switched off. Thus, among those complying with eligibility for tutoring, receiving tutoring alone increases the average grade by 2.58 points. However, this impact is far from being statistically significant at any conventional level, as the p-value (based on heteroscedasticity-robust standard errors) is equal to 55%.

The estimate of β_B suggests that among those who either (i) comply with their eligibility for financial incentives (self-compliers) *or* (ii) refuse the financial incentive when both treatments are available (cross-defiers), signing up for the financial incentive alone has a positive effect of 3.15 points. This effect is statistically significant at the 1% level. In contrast, the estimate of the interaction term β_{AB} is small and statistically insignificant. Nevertheless, as Theorem 3 shows, this term is not straightforward to interpret; the fact that it is close to zero does not, by itself, imply that their is no interference across the two treatments.

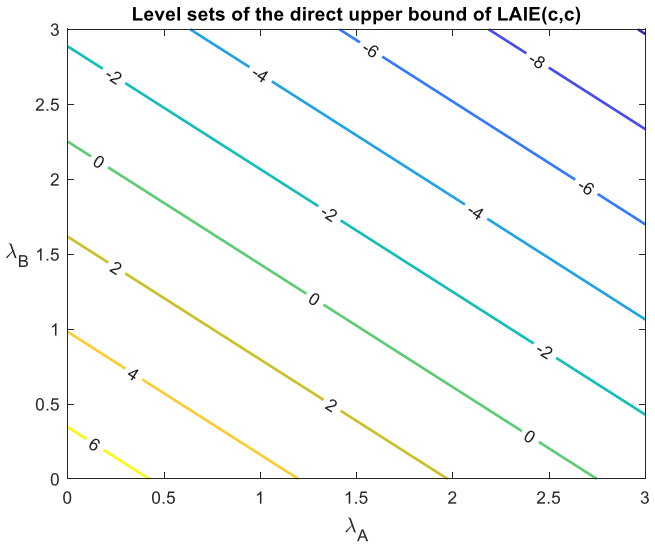

FIGURE 1 | Level sets of the upper bound in (16). $\lambda_A = ATE_{A|\overline{B}}(c,c)/ATE_{A|\overline{B}}(s \cup d, \cdot)$; $\lambda_2 = ATE_{B|\bar{A}}(c,c)/ATE_{B|\bar{A}}(\cdot, s \cup d)$.

We next illustrate how to employ the results in Section 4.3 to learn about the LAIE. We will ignore standard errors and treat all point estimates as if they were probability limits. We maintain Assumption 4 throughout, but impose type restrictions only as indicated.

5.2 | Direct Bounds on the Joint Effect and LAIE

We start by applying Theorem 4. Using the estimates from Tables S4.1 and S4.2, we obtain

$$U_{00}(c,c) = \frac{62.83}{0.49} - 70.55 \times \frac{0.19}{0.49} = 100.87 \text{ and}$$
$$L_{00}(c,c) = U_{00}(c,c) - 64.83 \times \frac{0.31}{0.49} = 59.85.$$

Given that $E[Y(11)|(c,c)] = 66.94$, this yields $ATE_{AB}(c,c) \in [-33.93, 7.09]$. Clearly, U_{00} could be replaced by 100, but it may also be replaced by 66.94 under the additional assumption that $Y(11) \geq Y(00)$, that is, that taking the two treatments jointly cannot hurt anybody's GPA. This improves the bounds for the joint effect to the reasonably tight interval [0,7.09] without requiring any type restrictions.

Heuristic Analysis

As suggested in Section 4.3, we can parameterize the standalone effects of D_A and D_B in the (c,c) subgroup as $ATE_{A|\overline{B}}(c,c) = \lambda_A ATE_{A|\overline{B}}(s \cup d, \cdot) = 2.58\lambda_A$ and $ATE_{B|\bar{A}}(c,c) = \lambda_B ATE_{B|\bar{A}}(\cdot, s \cup d) = 3.15\lambda_B$ for some multipliers $\lambda_A, \lambda_B \geq 0$. Equation (15) and the tightened bound on the joint effect then gives

$$-2.58\lambda_A - 3.15\lambda_B \leq LAIE(c,c) \leq 7.09 - 2.58\lambda_A - 3.15\lambda_B. \tag{16}$$

Figure 1 depicts the level sets of the upper bound in (16) for $\lambda_A, \lambda_B \in [0,3]$. Any combination (λ_A, λ_B) that lies above (i.e., northeast of) the zero line implies a negative LAIE, while for combinations below this line the sign of the interaction effect is unidentified. For example, if the average effect of D_A and D_B is about 25% larger among (c,c) types than among $(s \cup d, \cdot)$ and $(\cdot, s \cup d)$ types, respectively, then LAIE(c,c) is negative.

Formal Analysis

Adopting the restriction that there are no $(d,\cdot)$ and $(\cdot,j)$ pairs and computing the bounds in Theorem 5 yields $E[Y(10)|(c,c)] \in [37.28, 80.14]$ and $E[Y(01)|(c,c)] \in [59.38, 83.87]$. Even with the improved upper bound for $E[Y(00)|(c,c)]$, the implied LAIE lies in the interval $[-37.21, 37.22]$, which is very wide. If we impose the additional assumption that $Y(11) \geq \max\{Y(10), Y(01)\}$, then $E[Y(11)|(c,c)] = 66.94$ may be used as a tightened upper bound for $E[Y(10)|(c,c)]$ and $E[Y(01)|(c,c)]$ as well.[9] This shrinks the bound on the local interaction effect to $[-7.09, 37.22]$, but the sign remains unidentified.

5.3 | Indirect Bounds on LAIE

We impose the auxiliary condition $P(d,\cdot) = 0$ (see Section 5.1), but for the time being allow for joint compliers in the financial incentive treatment arm ($P(\cdot,j) > 0$). The expression for the interaction coefficient β_{AB} stated in Theorem 3 simplifies, and the local average interaction effect for (c,c) pairs can be expressed as

$$LAIE(c,c) = \beta_{AB} + \frac{P(j,\cdot)}{P(c,c)} ATE_{A|\overline{B}}(s,\cdot) + \frac{P(\cdot,j) - P(\cdot,d)}{P(c,c)} ATE_{B|\bar{A}}(\cdot, s \cup d) \tag{17}$$

$$- \frac{P(j,\cdot)}{P(c,c)} ATE_{A|\overline{B}}(j,\cdot) + \frac{P(\cdot,d)}{P(c,c)} ATE_{B|\bar{A}}(\cdot,d) - \frac{P(\cdot,j)}{P(c,c)} ATE_{B|\bar{A}}(\cdot,j), \tag{18}$$

where both average treatment effects under (17) are identified along with their the probability weights. By contrast, the average treatment effects under (18) are not identified and $P(\cdot,j)$ and $P(\cdot,d)$ are also not identified separately. For the identified quantities, we can substitute the point estimates from Tables 2–4 into (17) and (18) to obtain

$$LAIE(c,c) = 1.02 - 0.43 ATE_{A|\overline{B}}(j,\cdot) + \frac{0.12 + P(\cdot,j)}{0.49} ATE_{B|\bar{A}}(\cdot,d) - \frac{P(\cdot,j)}{0.49} ATE_{B|\bar{A}}(\cdot,j) \tag{19}$$

Just as in case of the direct bounds, we can proceed in two ways.

Heuristic Analysis

As suggested in Section 4.3, we use the identified local average treatment effects $ATE_{A|\overline{B}}(s,\cdot)$ and $ATE_{B|\bar{A}}(\cdot, s \cup d)$ as reference

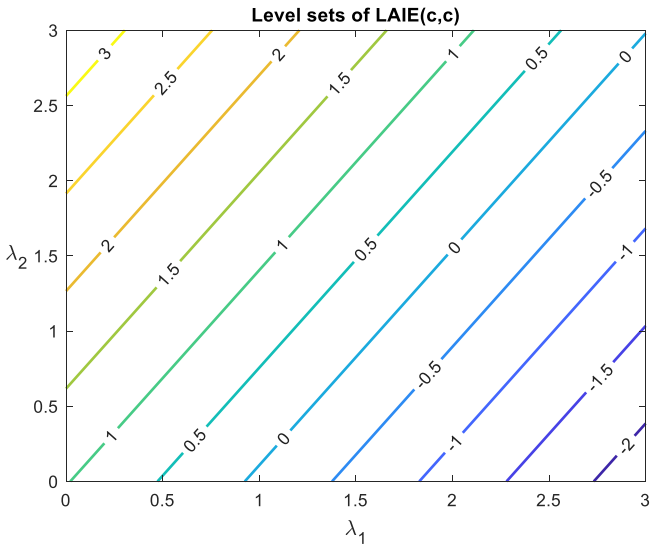

FIGURE 2 | Level sets of $LAIE(c,c)$ as a function of λ_1 and λ_2 when $P(\cdot,j)=0$. $\lambda_1 = ATE_{A|\overline{B}}(j,\cdot)/ATE_{A|\overline{B}}(s,\cdot)$; $\lambda_2 = ATE_{B|\bar{A}}(\cdot,d)/ATE_{B|\bar{A}}(\cdot,s\cup d)$. No joint compliers with respect to the financial incentive treatment.

points, and write

$$ATE_{A|\overline{B}}(j,\cdot) = \lambda_1 ATE_{A|\overline{B}}(s,\cdot),$$
$$ATE_{B|\bar{A}}(\cdot,d) = \lambda_2 ATE_{B|\bar{A}}(\cdot,s\cup d) \text{ and}$$
$$ATE_{B|\bar{A}}(\cdot,j) = \lambda_3 ATE_{B|\bar{A}}(\cdot,s\cup d),$$

where the λ_i, $i=1,2,3$ are scalar multipliers. Using the identified values of $ATE_{A|\overline{B}}(s,\cdot)$ and $ATE_{B|\bar{A}}(\cdot,s\cup d)$, and substituting into (19) gives

$$LAIE(c,c) = 1.02 - 1.11\lambda_1 + 6.43[0.12 + P(\cdot,j)]\lambda_2 - 6.43P(\cdot,j)\lambda_3. \quad (20)$$

If it is hypothesized that all the λ_i fall into, say, the interval [0,3], then (20) gives the following bounds on the interaction effect:

$$-2.30 - 19.29P(\cdot,j) \le ATE_{A|B}(c,c)$$
$$-ATE_{A|\overline{B}}(c,c) \le 3.34 + 19.29P(\cdot,j).$$

Clearly, the bounds are the tightest when $P(\cdot,j)=0$; that is, there are no joint compliers with respect to the financial incentive treatment. The sign of the interaction effect is not identified even in this case, which is not too surprising given that β_{AB} is so close to zero.

If one is willing to work with the assumption that $P(\cdot,j)=0$, expression (20) reduces to a function of λ_1 and λ_2 only, and it becomes more straightforward to evaluate $LAIE(c,c)$ in various hypothetical scenarios. In particular, Figure 2 shows the isoquants (level sets) of $LAIE(c,c)$ as a function of λ_1 and λ_2. From this graph one can read off the (λ_1,λ_2) pairs that are consistent with, say, a negative interaction between the treatments.

Formal Analysis

Alternatively, one can bound the unknown treatment effects in (19) by computing the Manski-type bounds stated in Theorem S1.1, while also imposing $P(\cdot,j)=0$. This yields, after replacing any uninformative bounds with trivial ones, $ATE_{A|\overline{B}}(j,\cdot) \in [0,43.88]$ and $ATE_{A|\overline{B}}(\cdot,d) \in [0,100]$. Combining these bounds with Equation (19) gives $LAIE(c,c) \in [-17.85,25.51]$, which is rather too loose to be useful. One can intersect this interval with the formal direct bounds for LAIE, but the sign remains unidentified. Nevertheless, these Manski-type bounds can still be informative in other applications.

6 | Conclusion

We study randomized experiments (or quasi-experiments) in which the experimental units are potentially exposed to one of two different treatments, both, or none. Compliance with the intended treatment assignments, described by two binary instruments, is allowed to be endogenous. Our setup allows for the presence of compliance types that, to our knowledge, have not been considered in the literature, but are needed to accommodate some applications. In particular, there can be individuals in the population for whom the presence of Z_B (or, resp. Z_A) represents a *negative* incentive to take treatment D_A (or, resp. D_B); we call this type cross-defiers. At the same time, we allow for joint compliers as well—a type that reacts positively to the partner instrument and ultimately takes a given treatment whenever both instruments are present.

We develop the causal interpretation of three IV estimands in our framework. The price of generality is that some of the identification results are weak in the sense that interesting causal parameters are inextricably tied up with terms arising from treatment effect heterogeneity, and auxiliary conditions are needed to obtain more useful interpretations. Alternatively, we provide partial identification results with the goal of bounding the interaction effect between the two treatments, which is frequently of interest in applications.

A clear advantage of the general approach is that one does not need to pre-commit to a theoretical framework that does not quite fit the data, and any further auxiliary conditions can be tailored to the application at hand. (For example, Blackwell 2017, needs to drop a small set of data points because they violate the treatment exclusion restriction.) Our empirical application, which analyzes a program randomly offering tutoring services (treatment A) and financial incentives (treatment B) to female college students, illustrates the advantages of starting from a general interpretative framework as well as the use of our partial identification results.

Acknowledgments

We thank Aniko Biro, Matias Cattaneo, Noemi Kreif, Attila Lindner, Laszlo Matyas, Timea Molnar, and three anonymous referees for useful comments. All errors are our responsibility. An earlier version of the paper was circulated under the title "Treatment Effect Analysis for Pairs with Endogenous Treatment Takeup."

Data Availability Statement

The data and computer code (in R language) necessary to replicate the results in our empirical application are available under the link https://doi.org/10.15456/jae.2025024.0714281503.

Endnotes

[1] In Section 2, we provide a brief literature review and position our paper more carefully relative to the most relevant subset of papers.

[2] One-sided noncompliance means that the treatment cannot be accessed without receiving the instrument.

[3] The basic features of our framework were originally developed in the MA thesis of Kormos (2018).

[4] He states a formula in the Supplemental Appendix and only notes that it does not have a direct causal interpretation.

[5] Defiance is only partial in the sense that the individual does not necessarily act against $Z_B = 0$, but adding this idea to the moniker would be tedious. An alternative label might be "deserters."

[6] A zero coefficient means that takeup is consistent with the treatment exclusion restriction, that is, only the own instrument matters.

[7] Suppose that individuals participate in a study on the health benefits of physical exercise. Specifically, there are two treatments: running (D_A) and swimming (D_B). The instrument Z_A is a seminar on the health benefits of running and Z_B a seminar on the benefits of swimming. A person who cannot swim will be a never taker with respect to D_B. Nevertheless, it is conceivable that for the same person $D_A(10) = 0$ but $D_A(11) = 1$. This means that a single lecture is not sufficient to convince this person to take up running but after hearing more about the health benefits of exercise, he eventually decides to do so.

[8] This result then holds even without one-sided non-compliance.

[9] The condition $Y(11) \geq \max\{Y(10), Y(01)\}$ implies that, for any individual, the joint effect is (weakly) larger than the standalone effect of each treatment. This assumption is rather strong but it still allows for the interaction effect to be potentially negative; see Equation (15).

References

Angrist, J. D., D. Lang, and P. Oreopoulos. 2009. "Incentives and Services for College Achievement: Evidence From a Randomized Trial." *American Economic Journal: Applied Economics* 1: 136–63.

Behaghel, L., B. Crépon, and M. Gurgand. 2013. "Robustness of the Encouragement Design in a Two-Treatment Randomized Control Trial." IZA Discussion Paper No 7447.

Bhuller, M., and H. Sigstad. 2024. "2SLS With Multiple Treatments." *Journal of Econometrics* 242: 105785.

Blackwell, M. 2017. "Instrumental Variable Methods for Conditional Effects and Causal Interaction in Voter Mobilization Experiments." *Journal of the American Statistical Association* 112: 590–599.

Ferracci, M., G. Jolivet, and G. J. van den Berg. 2014. "Evidence of Treatment Spillovers Within Markets." *Review of Economics and Statistics* 96: 812–823.

Fisher, R. A. 1925. *Statistical Methods for Research Workers*. Oliver and Boyd: Edinburgh.

Goff, L. 2022. "A Vector Monotonicity Assumption for Multiple Instruments." working paper arXiv:2009.00553.

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

Hong, G., and S. W. Raudenbush. 2006. "Evaluating Kindergarten Retention Policy." *Journal of the American Statistical Association* 101: 901–910.

Huber, M., and A. Steinmayr. 2021. "A Framework for Separating Individual-Level Treatment Effects From Spillover Effects." *Journal of Business & Economic Statistics* 39: 422–436.

Hudgens, M. G., and M. E. Halloran. 2008. "Toward Causal Inference With Interference." *Journal of the American Statistical Association* 103: 832–842.

Imai, K., Z. Jiang, and A. Malani. 2021. "Causal Inference With Interference and Noncompliance in Two-Stage Randomized Experiments." *Journal of the American Statistical Association* 116: 632–644.

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

Imbens, G. W., and C. F. Manski. 2004. "Confidence Intervals for Partially Identified Parameters." *Econometrica* 72: 1845–1857.

Kang, H., and G. Imbens. 2016. "Peer Encouragement Designs in Causal Inference With Partial Interference and Identification of Local Average Network Effects." Working Paper, arXiv.1609.04464.

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

Kormos, M. 2018. "Paired 2×2 Factorial Design for Treatment Effect Identification and Estimation in the Presence of Paired Interference and Noncompliance." (MA thesis), Central European University.

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

Manski, C. F. 1989. "The Anatomy of the Selection Problem." *Journal of Human Resources* 24: 343–360.

Manski, C. F. 1990. "Nonparametric Bounds on Treatment Effects." *American Economic Review* 80: 319–323.

Manski, C. F. 1997. "Monotone Treatment Response." *Econometrica* 65: 1311–1334.

Martínez-Iriarte, J. 2021. *Sensitivity Analysis in Unconditional Quantile Effects Working Paper*. University of California, Santa Cruz.

Masten, M. A., and A. Poirier. 2020. "Inference on Breakdown Frontiers." *Quantitative Economics* 11: 41–111.

Mogstad, M., A. Torgovitsky, and C. R. Walters. 2020. "Policy Evaluation With Multiple Instrumental Variables." NBER Working Paper 27546.

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

Rubin, D. B. 1978. "Bayesian Inference for Causal Effects." *Annals of Statistics* 6: 34–58.

Sobel, M. E. 2006. "What Do Randomized Studies of Housing Mobility Demonstrate? Causal Inference in the Face of Interference." *Journal of the American Statistical Association* 101: 1398–1407.

Vazquez-Bare, G. 2022. "Causal Spillover Effects Using Instrumental Variables." *Journal of the American Statistical Association*, forthcoming. https://doi.org/10.1080/01621459.2021.2021920.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.