# Treatment Effect Analysis for Pairs with Endogenous Treatment Takeup*

Mate Kormos†

Tinbergen Institute, Amsterdam

Robert P. Lieli‡

Central European University, Budapest.

February 14, 2019

**Abstract**

We study causal inference in a setting in which units consisting of *pairs* of individuals (such as married couples) are assigned randomly to one of four categories: treatment for pair member $A$ alone, a potentially different treatment for pair member $B$ alone, joint treatment, or no treatment. Allowing for endogenous non-compliance, coordinated treatment take-up, as well as interference across pair members, we derive the causal interpretation of various instrumental variable estimands. While the local average treatment effect (LATE) parameters associated with applying the two treatments in isolation are identified under appropriate monotonicity conditions, coordinated treatment takeup makes it difficult to separate the interaction between the two treatments from treatment effect heterogeneity. The stated identification results extend the literature on causal inference in settings where the stable unit treatment value assumption (SUTVA) does not hold.

*Keywords:* causal inference, interference, instrumental variables, non-compliance

*JEL codes:* C22, C26, C90

*We thank Laszlo Matyas for useful comments. All errors are our responsibility.

†Email: mate.kormos33@gmail.com

‡Email: lielir@ceu.edu

# 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. Comparing the average treatment and control outcomes then gives a consistent estimate of the average treatment effect across units. 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 critically examine these assumptions, and to extend the analysis of experimental data to more complicated situations such as the following.

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 heterogeneity in the anticipated outcomes under treatment vs. no treatment), then the average difference between treated and non-treated values does not represent the treatment effect alone but selection effects as well. In this case the original random assignment is better regarded as an instrumental variable (IV), and as shown in an influential paper by Imbens and Angrist (1994), one can use it to consistently estimate the average treatment effect for compliers—the local average treatment effect (LATE).

Second, randomization itself does not ensure that the treatment status of an individual unit 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. For example, Sobel (2006) describes an experiment in which randomly selected families from severely disadvantaged urban neighborhoods are provided vouchers to move. Whether or not a family actually moves, and any other subsequent outcome, may well depend on whether other families in their social circle decide to do so. This naturally complicates the causal interpretation of the difference between the treated and non-treated mean outcomes.

Third, there are experimental setups in which units are potentially subject to multiple, but not mutually exclusive, treatments that may interact with each other. For example, Blackwell (2017) studies a "get out the vote" field experiment in which households are

"randomly assigned to receive telephone canvassing, in-person canvassing, both, or neither [...]." A major goal of the analysis is to identify to what extent the two treatments reinforce or cancel out each other.

In this paper we study causal inference in an experimental or quasi-experimental setup that extends the basic Rubin causal model in all three directions mentioned above. The essential features of this framework are the following:

(i) The population is a set of pairs comprised of distinguishable members $A$ and $B$. There are two binary treatments, also labeled as $A$ and $B$, accessible to member $A$ and member $B$, respectively. The outcome of interest may be associated with member $A$ alone, member $B$ alone, or the pair itself.

(ii) There are two binary instruments, one targeted at member $A$ and one at member $B$, representing randomized assignment to the corresponding treatment or some exogenous incentive to take the treatment. Nevertheless, compliance is imperfect (endogenous), including possible coordination across pair members in treatment take-up, which is perhaps the most novel feature of our setup.

(iii) There is interference within pairs in that the potential outcomes of a pair member depend not only on their own treatment status but also on their partner's. There is no interference across pairs. In keeping with the philosophy of the Rubin causal model, individual level treatment effects are heterogeneous, and allow for arbitrary interactions between the two treatments.

Some examples will help fix ideas.

**Example 1:** Consider a group of married couples where one member suffers from depression and the other does not. Of interest are the effects of two binary treatments: (i) an antidepressant medication for the depressed spouse, and (ii) an educational program about depression for the healthy spouse. The dependent variable may measure the severity of the depression symptoms. While both treatments are likely to have an effect on their own, they may also interact—the medication might be more effective if accompanied by behavioral

adjustments on the partner's part. Moreover, even if the initial treatment assignments are random, the actual takeup decision is generally endogenous (dependent on the anticipated gains from treatment), and implicitly or explicitly coordinated across spouses. ∎

**Example 2:** Consider a population of working couples with a newborn baby. The two treatments in question are (i) an extended parental leave for the mother and (ii) an extended parental leave for the father. The outcome of interest is, say, a child development indicator. Whether both, none, or one of the parents go on leave, and in the last case the gender of the parent on leave, may matter for the outcome. One might design an experiment in which parents are offered a randomized incentive to go on leave together or separately. Still, the ultimate takeup pattern is likely to be coordinated, and dependent on unobserved factors that also affect the outcome. ∎

The question we ask is what type of causal effects can be identified from an experiment that uses a two-by-two factorial design to randomly assign each pair to one of four categories: no treatment, joint treatment, treatment for the first pair member only, treatment for the second pair member only. Alternatively, in quasi-experimental settings with observational data, explicit randomization is replaced by some type of exogenous variation in treatment assignment, e.g., the presence of an exogenous incentive to choose some of these categories over the others. Given these data, we study the causal interpretation of estimands defined by the probability limit of various IV estimators constructed over the full sample of available pairs or suitable subsamples.

Establishing the causal interpretation of these estimands is of course made challenging by endogenous non-compliance and interference among pair members. To facilitate identification we introduce auxiliary assumptions; notably, we assume that the randomized instrument satisfies one-sided non-compliance and other monotonicity conditions. As a result, each pair member belongs to one of three different types: *self-compliers* take up their treatment if and only if their own instrument is "turned on"; *joint compliers* take their treatment if and only if both instruments are on, and *never takers* do not take the treatment for any instrument configuration. Thus, there are nine possible types of pairs in theory, but we rule out two by assumption, analogously to the "no defiers" postulate in a standard IV setting.

We state identification results for various IV estimation strategies. We show that the LATE parameter associated with, say, treatment $A$ applied in isolation can be identified by a pair-level IV regression of the outcome on the treatment indicator in the subsample of pairs where member $B$ is "randomized out." This parameter is the average effect of treatment $A$ among pairs where member $A$ is a self-complier, or, equivalently, the average effect among pairs where member $A$ actually takes the treatment but member $B$ is randomized out of treatment $B$. The scope of this result is enhanced by the fact that it holds regardless of whether the outcome is associated with member $A$ alone, member $B$ alone, or the pair. In other words, the spillover effect on member $B$ of applying treatment $A$ to member $A$ is also identified.

The results are less positive if one is also interested in identifying interaction effects between the two treatments. To this end, we consider an IV regression of the outcome on the two treatment dummies and their product. With possible coordination in treatment takeup, the coefficient on the interaction term does not cleanly identify a local average interaction effect, i.e., the difference between the effect of treatment $A$ with $B$ turned on vs. off for a group of pairs. Instead, the interaction effect among complier-complier pairs (self or joint) is confounded by terms that measure the difference in the average effect of a given treatment among self-compliers vs. joint compliers. Another way to gauge interaction is to compare the LATE of, say, treatment $A$ with an IV regression of the outcome on the indicator of treatment $A$ in the subsample of pairs where member $B$ is "randomized in." The resulting IV estimand is a weighted average of various kinds of treatment effect parameters for pairs comprised of different combinations of types. Interaction effects are again confounded with heterogeneity: even if the effects of the two treatments do not depend on each other at all, this estimand may differ from LATE simply because treatment $A$ has, on average, a different effect among pairs with different compliance profiles.

Of course, we are not the first to address issues of interference in treatment effect estimation from experimental or observational data; see VanderWeele et al. (2014) and Athey and Imbens (2016) for reviews of the extant statistics and econometrics literature. Nevertheless, to our knowledge our paper is the first to combine endogenous non-compliance with inter-

acting treatments as well as coordinated treatment take-up. Here we briefly review some of the more closely related papers that our framework builds on.

Sobel (2006) studies experiments in which the intended treatment assignment is random, but units can freely coordinate on their takeup decisions. He shows how the difference between the average outcome in the assigned vs. unassigned sample (the traditional "intention to treat" effect) relates to appropriate causal estimands. Nevertheless, his framework features non-interacting treatments, and he focuses on randomization within a single (large) target group. By contrast, Halloran and Hudgens (2008) consider a treatment (e.g., vaccination) to be applied to individuals organized into several groups (e.g., villages). Treatment is assigned through a two-tiered randomization procedure (across and within groups), and individual treatment effects potentially depend on how many others are treated in the same group. A key difference between the setup in *ibid.* and ours is that they assume perfect compliance with the randomized assignment, which rules out any endogeneity (including coordination) in treatment takeup and facilitates identification.

The paper that is the closest to ours is Blackwell (2017). As mentioned above, *ibid.* considers two potentially interacting treatments (phone and in-person canvassing) targeted at a single unit (a household). While the setup allows for endogeneous non-compliance, the actual takeup decision is restricted similarly to Imbens and Kang (2016). Specifically, compliance with one treatment assignment is assumed to be unaffected by the other assignment (the "treatment exclusion restriction"). This is a strong assumption; for example, it means that answering a randomly assigned "get-out-the-vote" phone call does not make one any more or less likely to answer the door to engage with an in-person canvasser. By contrast, pair members in our setup are allowed to make the two compliance decisions contingent on each other. As individual units can always be represented as pairs with identical members, our framework formally nests Blackwell's, and most of our identification results can be regarded as generalizations of *ibid.* to the coordinated treatment take-up case.

The rest of the paper is organized as follows. Section 2 presents a potential outcome framework with endogenous compliance for pairs. We state and discuss our identification results in Section 3. Section 4 concludes. Proofs are collected in the Appendix.

# 2 A potential outcome framework for pairs

## 2.1 Variable definitions

The population consists of ordered pairs of individuals (e.g., married couples), and pair members are assumed to be distinguishable. Ordering within a pair is based on some observed characteristic (e.g., husband/wife or depressed/healthy spouse). We will refer to the first member of a pair as member $A$ and the second as member $B$. Each pair in the population is potentially exposed to two different treatments—one targeted at member $A$ and one targeted at member $B$. We will denote the corresponding treatment status indicators as $D_A$ and $D_B$, respectively.

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 $D_B = 0$, i.e., 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

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

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 for (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.

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

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

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 (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) $P(D_A(10) = 1) > 0$, $P(D_B(01) = 1) > 0$, and $0 < P(Z_A = Z_B) < 1$.

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 construction in an experimental setting, where intended treatment assignments are explicitly randomized. Part (iii) states that for pair members $A$ and $B$, the fraction of those who comply with their own assignment regardless of their partner's is non-zero. Furthermore, the two instruments are not perfectly correlated, so $P(Z_A = i, Z_B = j) > 0$ for all choices $i, j \in \{0,1\}$.

We will impose further monotonicity assumptions on the potential treatment indicators in Section 2.3.

## 2.2 Parameters of interest

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

- $ATE_{A|\bar{B}}(\mathcal{P}) = E[Y(10) - Y(00) | \mathcal{P}]$ denotes the average effect of applying treatment $A$ alone to the subpopulation $\mathcal{P}$ relative to applying no treatment at all; in other words, this is the average effect of treatment $A$ conditional on treatment $B$ being "turned off."

- $ATE_{A|B}(\mathcal{P}) = E[Y(11) - Y(01)|\,\mathcal{P}]$ 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 $ATE_{A|\bar{B}}(\mathcal{P})$ and $ATE_{A|B}(\mathcal{P})$ are called local average conditional effects, or LACEs, by Blackwell (2017), while $ATE_{AB}(\mathcal{P})$ is called the local average joint effect (LAJE). For a given $\mathcal{P}$, the local average joint effect is the sum of the two conditional effects. The difference of the two conditional effects, $ATE_{A|B}(\mathcal{P}) - ATE_{A|\bar{B}}(\mathcal{P})$, measures the interaction between the two treatments in group $\mathcal{P}$ and is termed the local average interaction effect (LAIE) by *ibid.* If this quantity 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.

Given the potential interference across pair 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 $ATE_{A|\bar{B}}(\mathcal{P})$ and $ATE_{A|B}(\mathcal{P})$ measure what is called the direct effect of treatment $A$ by Halloran and Hudgens (2008). On the other hand, if $Y$ is associated with member $B$ alone, then $ATE_{A|\bar{B}}(\mathcal{P})$ 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.

## 2.3 Relevant subpopulations (types of pairs)

The setup presented in Section 2.1 assigns four potential treatment indicators to each pair member, corresponding to the four possible incentive schemes $(Z_A, Z_B)$ presented to the pair. Without any restrictions on the self-selection process into treatment, the possible configurations of these 8 variables partition the population of pairs into $2^8 = 256$ different

types in terms of their response to incentives. At this level of generality a couple of regression coefficients can hardly be an informative summary of the various average treatment effects across types. Therefore, we will impose empirically plausible assumptions on the potential treatment indicators, which dramatically reduces the number of relevant subpopulations, and facilitates the interpretation of IV regressions.

ASSUMPTION 2 (i) One-sided non-compliance with own instrument: $D_A(0,z) = 0$ and $D_B(z,0) = 0$ for $z \in \{0,1\}$. (ii) Monotonicity in partner's instrument: $D_A(z,0) \leq D_A(z,1)$ and $D_B(0,z) \leq D_B(1,z)$ for $z \in \{0,1\}$.

Assumption 2(i) states that neither member of the pair has access to their own treatment unless they have been "randomized in," i.e., the value of their own instrument is 1. Whether or not this assumption is reasonable depends on the institutional setting and details of the underlying experiment. In general, one-sided non-compliance presumes that the experimenter is able to exclude individuals from treatment and the treatment is hard to substitute for.

Part (ii) of Assumption 2 is implied by part (i) for $z = 0$, and for $z = 1$ it states that the presence of the partner's incentive (or the fact that the partner has been "randomized in") cannot reverse one's decision to take the treatment. However, the partner's incentive can still positively affect one's own treatment take-up. Hence, Assumption 2(ii) is more general than the personalized treatment assumption in Kang and Imbens (2016) and Blackwell (2017), which requires $D_A(z,0) = D_A(z,1)$ and $D_B(0,z) = D_B(1,z)$ for $z \in \{0,1\}$. Whether Assumption 2(ii) is plausible depends primarily on the specifics of the underlying experiment.[1] While it is not testable at the individual level, it has testable "aggregate" implications. Specifically, if one runs an OLS regression of, say, $D_A$ on $Z_B$ and a constant in the $Z_A = 1$ subsample, then the coefficient on $Z_B$ should be positive. A zero (insignificant) coefficient is consistent with the pearsonalized treatment assumption.

Assumption 2 is rather powerful in that it greatly reduces the number of relevant "com-

[1]In some applications monotonicity might hold in the other direction, i.e., $D_A(z,0) \geq D_A(z,1)$ and $D_B(z,0) \geq D_B(z,1)$. This assumption gives rise to a complementary theory, which is not discussed in this version of the paper.

pliance types" a pair member can belong to. The following definition presents the three remaining possibilities.

DEFINITION 1 Under Assumption 2, member $A$ of a pair $(A, B)$ is called a

- *self-complier* if $D_A(10) = 1$;
- *joint complier* if $D_A(10) = 0$ and $D_A(11) = 1$;
- *never taker* if $D_A(10) = 0$ and $D_A(11) = 0$.

Furthermore, a pair member is a *complier* if they are either a self-complier or joint-complier. The set of self-compliers, joint compliers, never takers and compliers is abbreviated as $s$, $j$, $n$ and $c$, respectively.

**Remarks**

1. The corresponding definitions for member $B$ can be stated in a similar way; these are omitted for brevity.[2]
2. A self-complier's treatment status is determined solely by the value of their own instrument. To see this, note that for member $A$, $D_A(00) = D_A(01) = 0$ by one-sided non-compliance and $D_A(10) = D_A(11) = 1$ by definition and monotonicity in $B$'s instrument. Therefore, $D_A = Z_A$.
3. 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. For a type $A$ joint complier this means $D_A = Z_A Z_B$.
4. Finally, a never taker cannot be induced to take the treatment by any instrument configuration.

Given that each pair member is a self-complier, joint complier or never taker, there are nine conceivable types of pairs:

$$\{(s,s),(s,j),(s,n),(j,s),(j,j),(j,n),(n,s),(n,j),(n,n)\},$$

[2]Replace the subscript $A$ with $B$ on the potential outcome variables and interchange the two arguments.

where, say, $(s, j)$ is the set of pairs where $A$ is a self-complier ($A \in s$) and $B$ is a joint complier ($B \in j$), etc. Similarly, the notation $(c, \cdot)$ stands for the set of pairs where member $A$ is a complier, etc. We denote the population proportion of the $(s, s), (s, j), \ldots$ pairs as $P(s, s)$, $P(s, j)$, etc.

Our last assumption removes two types of pairs from consideration.

ASSUMPTION 3 $P(n, j) = P(j, n) = 0$.

Assumption 3 states that if one pair member never takes the treatment, then the other will also disregard the experimenter's intention for that member. This scenario is plausible if pair members know each others' types and the incentive represented by the instrument is not transferable across members. For example, $Z_B$ could be a randomized monetary reward payable on actual treatment take-up by member $B$. If $B$ is a never taker, then $B$ will not receive the payment, and hence cannot share it with $A$. It is then reasonable for $A$ to disregard $B$'s instrument assignment in their own treatment decision.[3] Another way to interpret this assumption is that member $A$'s utility of taking taking treatment $A$ is affected by $Z_B$ only through member $B$'s actual treatment status $D_B$. Once member $A$ knows that $B$ is a never taker, it follows that $D_B = 0$, and the value of $Z_B$ is no longer relevant for $A$'s decision. Hence $A$ cannot be a joint complier. The same reasoning applies if one interchanges the role of $A$ and $B$.

Figure 1 illustrates the partitioning of the population of pairs into compliance types induced by Assumptions 2 and 3. The exact type of a given pair is generally unobserved as it depends on the pair's behavior in counterfactual scenarios. Nevertheless, partial identification is still possible; for example, if $D_A = D_B = 1$, then this pair is clearly not $(n, n)$. As a result, one can draw inferences about the relative frequency of certain types of pairs, or groups of types, from the observed propensity scores

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

The following lemma establishes the population proportion of certain types of pairs as a

[3] In some situations $B$ may be willing/able to compensate $A$ privately for being a never taker. However, such compensation is unlikely to be contingent on $Z_B$.

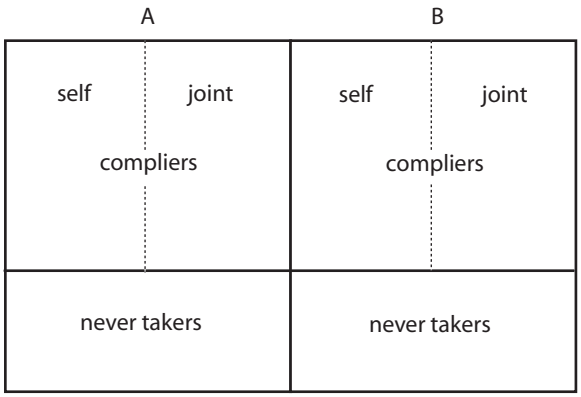


Figure 1: Possible types of pair members. Not all individual types are compatible with each other; never taker - joint complier pairings are ruled out.

function of the compliance probabilities $\pi$. These proportions will play the role of probability weights in subsequent identification results presented in Section 3.

**Lemma 1** *Suppose that Assumption 2 is satisfied. Then:*

$$P(s,\cdot) = \pi(10|10),\ P(\cdot,s) = \pi(01|01),\ P(c,n) = \pi(10|11),\ P(n,c) = \pi(01|11),\ \ \textit{and}$$
$$P(c,c) = \pi(11|11).$$

*Furthermore,*

$$P(c,\cdot) = P(c,c) + P(c,n) \ \ \textit{and} \ \ P(\cdot,c) = P(c,c) + P(n,c);$$
$$P(j,\cdot) = P(c,c) + P(c,n) - P(s,\cdot) \ \ \textit{and} \ \ P(\cdot,j) = P(c,c) + P(n,c) - P(\cdot,s).$$

*If, in addition, Assumption 3 holds, then*

$$P(c,s) = P(\cdot,s) - P(n,s) = P(\cdot,s) - P(n,c)$$
$$P(s,c) = P(s,\cdot) - P(s,n) = P(s,\cdot) - P(c,n)$$
$$P(c,j) = P(\cdot,j) = P(c,c) + P(n,c) - P(\cdot,s)$$
$$P(j,c) = P(j,\cdot) = P(c,c) + P(c,n) - P(s,\cdot).$$

**Remarks**

1. The proof of Lemma 1 is stated in Appendix A.

2. One cannot identify the relative frequency of all primitive types $(k, l) \in \{n, s, j\}^2$. As we argue in Appendix A, out of the 16 compliance probabilities $\pi(d_A, d_B|z_A, z_B)$, there are at most five that carry independent information about types under Assumption 2 (such a selection is shown in the first paragraph of Lemma 1). At the same time, there are nine conceivable pairs, which reduces to seven if Assumption 3 is imposed. This still means that six unknowns would need to be recovered from five linear restrictions.

# 3 The causal interpretation of three IV estimands

## 3.1 Estimators and estimands

Each pair in the target population is associated with a random 5-vector $(Y, D_A, D_B, Z_A, Z_B)$. The following assumption imposes an additional technical condition on the distribution of this vector and describes the sample data.

ASSUMPTION 4 (i) $E(|Y|^2) < \infty$. (ii) Let $\{(Y_i, D_{A,i}, D_{B,i}, Z_{A,i}, Z_{B,i})\}_{i=1}^n$ be independent and identically distributed draws from the population of pairs targeted by the treatments.

We study a number of IV estimators, constructed either over 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 $D_A$. Of interest is the estimated slope coefficient:

$$\hat{\delta}_{A0} = \frac{\sum_{i:Z_{B,i}=0}(Y_i - \bar{Y})(Z_{A,i} - \bar{Z}_A)}{\sum_{i:Z_{B,i}=0}(D_{A,i} - \bar{D}_A)(Z_{A,i} - \bar{Z}_A)},$$

where the upper bar denotes the sample mean in the $Z_B = 0$ subsample.

By Assumption 4 and standard arguments based on the law of large numbers, the probability limit of $\hat{\delta}_{A0}$ as $n \to \infty$ is given by the Wald estimand

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

As $\delta_{A0}$ derives from a pair-level regression, inference about this parameter does not require clustered standard errors as in Halloran and Hudgens (2008).

(ii) The IV regression described in point (i) above can also be implemented in the $Z_B = 1$ subsample. Denoting the slope coefficient by $\hat{\delta}_{A1}$, its probability limit is given by another Wald estimand

$$\delta_{A1} = \frac{E(Y|Z_A = 1, Z_B = 1) - E(Y|Z_A = 0, Z_B = 1)}{E(D_A|Z_A = 1, Z_B = 1) - E(D_A|Z_A = 0, Z_B = 1)}. \tag{2}$$

(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 $\ddot{Z} = (1, Z')'$. The IV estimator is a $4 \times 1$ vector $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_A, \hat{\beta}_B, \hat{\beta}_{AB})'$ given by

$$\hat{\beta} = \left( \sum_{i=1}^{n} \ddot{Z}_i \ddot{D}_i' \right)^{-1} \sum_{i=1}^{n} \ddot{Z}_i Y_i.$$

Again, taking the probability limit of $\hat{\beta}$ under Assumption 4 yields the estimand

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

## 3.2 Identification results

The following three theorems, the main results of this paper, state the causal interpretation of the estimands defined above.

**Theorem 1** *Under Assumptions 1 and 2, the Wald estimand (1) is equal to*

$$ATE_{A|\bar{B}}(s, \cdot).$$

**Remarks**

1. The proof of Theorem 1 follows the classic argument by Imbens and Angrist (1994) and is omitted for brevity.

2. Theorem 1 states that the Wald estimand identifies the average effect of treatment $A$ alone among pairs where member $A$ is a self-complier. In other words, member $A$ is a complier with respect to the instrument $Z_A$ alone, i.e., $ATE_{A|\bar{B}}(s,\cdot)$ is a usual LATE parameter.

3. It is worth recalling that $Y$ could be an outcome associated solely with member $B$. In this case the Wald estimand identifies the average *spillover* effect on member $B$ of a treatment applied to member $A$ — at least among pairs where $A$ is a self-complier.

4. There is of course a corresponding result for treatment $B$ which can be obtained by interchanging the subscripts $A$ and $B$ throughout expression (1) and Theorem 1, and replacing $(s,\cdot)$ with $(\cdot,s)$.

5. In standard settings with SUTVA, if a binary instrument satisfies one-sided non-compliance, the associated LATE parameter reduces to ATT (e.g., Angrist and Pischke 2008, Donald et al. 2014). Similarly, an alternative interpretation of $ATE_{A|\bar{B}}(s,\cdot)$ is that it gives the average effect of the treatment $A$ among all pairs where $A$ is treated and $Z_B = 0$.[4] This is an observed subset of the population.

The Wald estimand (1) has a straightforward interpretation because the condition $Z_B = 0$ completely rules out treatment participation by member $B$. By contrast, the estimand (2) conditions on $Z_B = 1$, which allows for a number of options in member $B$'s takeup decision. As a result, the causal interpretation of (2) is more complicated.

**Theorem 2** *Under Assumptions 1 through 3, the Wald estimand (2) is equal to*

$$ATE_{AB}(c,j)\frac{P(c,j)}{P(c,\cdot)} + ATE_{A|B}(c,s)\frac{P(c,s)}{P(c,\cdot)} + ATE_{A|\bar{B}}(c,n)\frac{P(c,n)}{P(c,\cdot)}. \tag{3}$$

[4]To see this, note that $D_A = D_A(11)Z_AZ_B + D_A(10)Z_A(1-Z_B)$ by one-sided non-compliance and therefore $D_A(10) = 1, Z_B = 0 \Leftrightarrow D_A = 1, Z_B = 0$. Letting $\Delta = Y(10) - Y(00)$, this implies

$$ATE_{A|\bar{B}}(s,\cdot) = E[\Delta|\, D_A(10) = 1] = E[\Delta|\, D_A(10) = 1, Z_B = 0] = E[\Delta|\, D_A = 1, Z_B = 0].$$

**Remarks**

1. The proof of Theorem 2 is given in Appendix B.

2. Theorem 2 states that the Wald estimand (2) can be interpreted as the weighted average of three causal effects: (i) the average effect of the *joint* treatment among pairs where $A$ is a complier (self or joint) and $B$ is a joint complier; (ii) the average effect of treatment $A$ with $B$ turned on among pairs where $A$ is a complier and $B$ is a self-complier; and (iii) the average effect of treatment $A$ with $B$ turned off among pairs where $A$ is a (self-)complier and $B$ is a never taker.

3. $P(c,\cdot) = P(c,j) + P(c,s) + P(c,n)$, i.e., the weights in (3) sum to one. Thus, the ratio $P(c,j)/P(c,\cdot)$ is the probability that member $B$ is a joint complier given that member $A$ complier, etc. Lemma 1 shows how these probability weights can be identified from the observed data.

4. Despite looking complicated, the result in Theorem 2 is very intuitive. Conditional on $Z_B = 1$, consider changing $Z_A$ from 0 to 1. In this case $(c,j)$ pairs will switch from no treatment at all to both treatments, contributing the first term in (3). 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.

5. Again, the interpretation of Theorem 2 is enriched by the fact that $Y$ can be an outcome associated with $A$ alone, $B$ alone, or the pair $(A,B)$.

6. The corresponding result for the IV regression involving treatment $B$ in the $Z_A = 1$ subsample can be obtained by interchanging the role of $A$ and $B$ in (2) and (3).

Finally, Theorem 3 states the causal interpretation of the components of $\beta$.

**Theorem 3** *Under Assumptions 1 through 3,*

$$\begin{aligned}
\beta_0 &= E[Y(00)] \\
\beta_A &= ATE_{A|\bar{B}}(s,\cdot) \\
\beta_B &= ATE_{B|\bar{A}}(\cdot,s) \\
\beta_{AB} &= ATE_{A|B}(c,c) - ATE_{A|\bar{B}}(c,c) && (4) \\
&+ \frac{P(j,\cdot)}{P(c,c)}[ATE_{A|\bar{B}}(j,\cdot) - ATE_{A|\bar{B}}(s,\cdot)] && (5) \\
&+ \frac{P(\cdot,j)}{P(c,c)}[ATE_{B|\bar{A}}(\cdot,j) - ATE_{B|\bar{A}}(\cdot,s)]. && (6)
\end{aligned}$$

**Remarks**

1. The proof of Theorem 3 is given in Appendix C.

2. The coefficients on the stand-alone treatment dummies are the same as in the split sample case, i.e., they identify the average effect of the treatment in question (with the other turned off) among pairs where the corresponding member is a self-complier.

3. The coefficient on the interaction term can be written in several forms and has a complex interpretation. In Theorem 3 we exhibit it as the sum of three different terms. Term (4) is the local average interaction effect between the two treatments among $(c,c)$ pairs. Term (5) compares the average effect of treatment $A$ alone (with treatment $B$ turned off) across two different subpopulations—pairs where member $A$ is a self complier vs. a joint complier. Term (6) has an analogous interpretation for treatment $B$.

4. Theorem 3 generalizes Theorem 2 by Blackwell (2017) in that it does not impose the treatment exclusion restriction (i.e., allows for coordinated treatment takeup decisions) or require the independence of the instruments. (Nevertheless, we do rely on one-sided non-compliance, which is not generally imposed by *ibid.*) The key difference between the two results is in the interpretation of the interaction coefficient. Under Blackwell's assumptions, there is no distinction between self-compliers and joint compliers, as all compliers are of the former type. Hence the terms (5) and (6) vanish, and the resulting

coefficient (4) identifies the interaction between the two treatments among $(s, s)$ pairs. (This is true regardless of whether one-sided non-compliance is imposed.) In contrast, allowing for coordination necessitates a distinction between the two types of compliers. As a result, the interaction coefficient no longer represents a pure interaction effect, but also whether the two treatments act differently (on average) among self-compliers and joint compliers. Isolating the interaction effect (4) is not possible without further assumptions.

5. The proof of Theorem 3 is based on relating the two-stage least squares formulation of the IV estimand $\beta$ to the reduced form regression, i.e., the regression of $Y$ on $\ddot{Z}$. In the reduced form regression the coefficients are intention to treat effects, i.e., they are closely related to the numerators of (1) and (2). The causal interpretation of these quantities derive from Theorems 1 and 2. On the other hand, the same coefficients can be thought of as the "product" of the first stage coefficients (from the regression of $D$ on $\ddot{Z}$) with the second stage coefficients (from the regression of $Y$ on the predicted values of $D$ and a constant). Of course, the second stage coefficients coincide with $\beta$, and hence the causal interpretation of $\beta$ follows from "dividing" the reduced form coefficients by the first stage coefficients.

# 4 Conclusion

We introduced a framework for analyzing data from randomized experiments (or quasi-experiments) in which the experimental units are pairs, and there are two different treatments targeted at the two pair members. In order to accommodate realistic applications, compliance with the intended treatment assignments is assumed to be endogenous, including possible coordination among pair members in treatment takeup. Furthermore, there is interference within pairs, i.e., the treatment targeted at, say, member $A$ may have spillover effects on the outcomes of member $B$ (and vice versa). While previous research has addressed some of these problems in isolation, to our knowledge our paper is the first to study treatment effect identification in a framework that combines all three issues (non-compliance,

coordination and interference).

We stated results for three IV estimators. First, we considered the Wald estimator of treatment $A$ in the subsample where the instrument for member $B$ was turned off. The probability limit of this estimator is the average effect of treatment $A$ among those pairs where member $A$ is a self-complier. Hence, this is the LATE of treatment $A$ applied in isolation in the standard sense. Second, we considered the Wald estimator of treatment $A$ in the subsample where the instrument for member $B$ was turned on. The interpretation of this estimand is more complex; it is the weighted average of three LATE-type parameters that characterize the effect of treatment $A$ alone or in conjunction with treatment $B$. In particular, the estimand combines (i) the average joint effect of the two treatments among $(c, j)$ pairs; (ii) the average effect of treatment $A$ in the presence of treatment $B$ among $(c, s)$ pairs; and (iii) the average effect of treatment $A$ alone among $(c, n)$ pairs. Finally, we considered the full-sample IV regression of the outcome on the two treatment dummies and their interaction, instrumented by the intended assignment for member $A$, member $B$, and their interaction. The regression coefficients on the stand-alone treatment dummies can again be interpreted as standard LATE parameters, while the coefficient on the interaction term is a more complicated weighted sum that depends (i) on the interaction between the two treatments among $(c, c)$ pairs, and (ii) on how different the average effects of treatment $A$ and $B$ are among joint compliers versus self-compliers.

A general conclusion that follows from these results is that if the treatment exclusion restriction does not apply, i.e., pair members can, in part, base their participation decision on what their partner is doing, then it becomes very hard to separate treatment interaction from treatment effect heterogeneity. If the former is of interest, one must either defend the "no coordination" (treatment exclusion) assumption credibly or introduce some other identifying assumption that allows one to separate the two of effects.

The development of an empirical application that illustrates the theoretical results presented in this paper is work in progress.

# Appendix

## A. Proof of Lemma 1

For $Z_A = Z_B = 0$, the four probabilities $\pi(d_A, d_B|\,00)$ are all zero or one by one-sided non-compliance, and hence uninformative about the relative frequency of types. For $Z_A = 1, Z_B = 0$, we have $\pi(01|\,10) = \pi(11|\,10) = 0$, but

$$\begin{aligned}\pi(10|\,10) &= P(D_A = 1, D_B = 0|\,Z_A = 1, Z_B = 0) = P[D_A(10) = 1, D_B(10) = 0|\,Z_A = 1, Z_B = 0] \\ &= P[D_A(10) = 1, D_B(10) = 0] = P[D_A(10) = 1] = P(s, \cdot),\end{aligned}$$

where the third equality uses Assumption 1, and the fourth uses Assumption 2. The probability $\pi(00|\,10)$ is simply the complement of $\pi(10|\,10)$ and does not carry independent information. A similar reasoning applies to the case $Z_A = 0, Z_B = 1$, yielding the result $\pi(01|\,01) = P(\cdot, s)$. Finally, for $Z_A = Z_B = 1$, any three of the four probabilities $\pi(d_A, d_B|\,11)$ are informative; the stated interpretation of $\pi(01|\,11)$, $\pi(10|\,11)$, and $\pi(11|\,11)$ can be verified by simple calculations as above, using the definition of the various types. The same goes for all the other results. In sum, there are only $0 + 1 + 1 + 3 = 5$ independently informative compliance probabilities, precluding the identification of $P(k, l)$ for all seven pairs $(k, l) \in \{n, s, j\}^2 \setminus \{(n, j), (j, n)\}$.

## B. Proof of Theorem 2

We will first analyze the denominator in (2). Conditional on the event $\{Z_A = 1, Z_B = 1\}$, we have $D_A = D_A(11)$, and by random assignment (Assumption 1(ii)),

$$E[D_A|Z_A = 1, Z_B = 1] = E[D_A(11)] = P(D_A(11) = 1) = P(c, \cdot),$$

where the last equality uses Definition 1. On the other hand, $E[D_A|Z_A = 0, Z_B = 1] = 0$ because $Z_A = 0$ implies $D_A = 0$ by Assumption 2(i). Therefore, the denominator of (2) is equal to $P(c, \cdot)$.

We will now turn to the numerator. Conditional on $\{Z_A = 1, Z_B = 1\}$ the observed outcome $Y$ is

$$\begin{aligned}&Y(11)D_A(11)D_B(11) + Y(10)D_A(11)[1 - D_B(11)] + Y(01)[1 - D_A(11)]D_B(11) \\ &+Y(00)[1 - D_A(11)][1 - D_B(11)]. \qquad (7)\end{aligned}$$

On the other hand, conditional on $\{Z_A = 0, Z_B = 1\}$, the observed outcome $Y$ is simply

$$Y(01)D_B(01) + Y(00)[1 - D_B(01)], \qquad (8)$$

as $D_A(01) = 0$ by Assumption 2(i). Then, by random assignment (Assumption 1(ii)),

$$E[Y|Z_A = 1, Z_B = 1] - E[Y|Z_A = 0, Z_B = 1] = E[(7) - (8)], \qquad (9)$$

where the expectation on the r.h.s. of (9) is with respect to the joint distribution of the random variables

$$Y(i,j), i,j \in \{0,1\}, D_A(11), D_B(01), D_B(11).$$

To calculate this expectation, we first condition on the eight mutually exclusive (and exhaustive) events defined by the possible values of the vector $(D_A(11), D_B(01), D_B(11))$. We then average these conditional expectations using the probability weights of the various events. The eight cases are summarized below:

| Event | Interpretation | Probability | (7)−(8) |
|---|---|---|---|
| $D_A(11)=1, D_B(01)=0, D_B(11)=1$ | $A \in c,\ B \in j$ | $P(c,j)$ | $Y(11)-Y(00)$ |
| $D_A(11)=1, D_B(01)=1, D_B(11)=1$ | $A \in c,\ B \in s$ | $P(c,s)$ | $Y(11)-Y(01)$ |
| $D_A(11)=1, D_B(01)=0, D_B(11)=0$ | $A \in c,\ B \in n$ | $P(c,n)=P(s,n)^*$ | $Y(10)-Y(00)$ |
| $D_A(11)=1, D_B(01)=1, D_B(11)=0$ | $\emptyset^{**}$ | 0 | $Y(10)-Y(01)$ |
| $D_A(11)=0, D_B(01)=0, D_B(11)=1$ | $A \in n,\ B \in j$ | $0^*$ | $Y(01)-Y(00)$ |
| $D_A(11)=0, D_B(01)=1, D_B(11)=1$ | $A \in n,\ B \in s$ | $P(n,s)$ | 0 |
| $D_A(11)=0, D_B(01)=0, D_B(11)=0$ | $A \in n,\ B \in n$ | $P(n,n)$ | 0 |
| $D_A(11)=0, D_B(01)=1, D_B(11)=0$ | $\emptyset^{**}$ | 0 | $Y(00)-Y(01)$ |

*Notes:* *By Assumption 3, $P(j,n)=P(n,j)=0$.

**This case is ruled out by Assumption 2(ii) (monotonicity in partner's instrument).

Clearly, only the first three cases contribute to the r.h.s. of (9), and they yield:

$$\begin{aligned}
&E[Y|Z_A=1,Z_B=1]-E[Y|Z_A=0,Z_B=1]\\
&=E[Y(11)-Y(00)|A\in c,B\in j]P(c,j)+E[Y(11)-Y(01)|A\in c,B\in s]P(c,s)\\
&+E[Y(10)-Y(00)|A\in c,B\in n]P(c,n)\\
&=ATE_{AB}(c,j)P(c,j)+ATE_{A|B}(c,s)P(c,s)+ATE_{A|\bar{B}}(c,n)P(c,n).
\end{aligned}$$

Combining this result with the fact that the denominator of (2) is equal to $P(c,c)$ completes the proof.

## C. Proof of Theorem 3

**First stage** The first stage of the IV estimator consists of regressing $D_A$, $D_B$ and $D_AD_B$ on $Z=(Z_A,Z_B,Z_AZ_B)'$ and a constant to obtain the linear projections (predicted values) $\hat{D}_A$, $\hat{D}_B$ and $\widehat{D_AD_B}$.

More specifically, write

$$\begin{aligned}
D_A &= \gamma_{A0} + \gamma_{AA} Z_A + \gamma_{AB} Z_B + \gamma_{A,AB} Z_A Z_B + U_A \\
&= \gamma_{A0} + \gamma_A' Z + U_A \\
D_B &= \gamma_{B0} + \gamma_{BA} Z_A + \gamma_{BB} Z_B + \gamma_{B,AB} Z_A Z_B + U_B \\
&= \gamma_{B0} + \gamma_B' Z + U_B \\
D_A D_B &= \gamma_{AB,0} + \gamma_{AB,A} Z_A + \gamma_{AB,B} Z_B + \gamma_{AB,AB} Z_A Z_B + U_{AB} \\
&= \gamma_{AB,0} + \gamma_{AB}' Z + U_{AB},
\end{aligned}$$

where $E[U_A] = E[U_B] = E[U_{AB}] = 0$, $E[U_A Z] = E[U_B Z] = E[U_{AB} Z] = 0$, and $\gamma_A' = (\gamma_{AA}, \gamma_{AB}, \gamma_{A,AB})$, etc. By standard linear regression theory, the coefficients from the projection of $D_A$ have the following interpretations:

$$\gamma_{A0} = E(D_A |\, Z_A = 0, Z_B = 0) \tag{10}$$

$$\gamma_{AA} = E(D_A |\, Z_A = 1, Z_B = 0) - E(D_A |\, Z_A = 0, Z_B = 0) \tag{11}$$

$$\gamma_{AB} = E(D_A |\, Z_A = 0, Z_B = 1) - E(D_A |\, Z_A = 0, Z_B = 0) \tag{12}$$

$$\begin{aligned}
\gamma_{A,AB} &= E(D_A |\, Z_A = 1, Z_B = 1) - E(D_A |\, Z_A = 0, Z_B = 1) \\
&- \big[E(D_A |\, Z_A = 1, Z_B = 0) - E(D_A |\, Z_A = 0, Z_B = 0)\big].
\end{aligned} \tag{13}$$

Using the IV assumptions 1 through 3, particularly independence and one-sided non-compliance w.r.t. one's own instrument, these coefficients reduce to:

$$\gamma_{A0} = E[D_A(00)] = 0 \tag{14}$$

$$\gamma_{AA} = E[D_A(10)] = P(s, \cdot) \tag{15}$$

$$\gamma_{AB} = E[D_A(01)] - E[D_A(00)] = 0 \tag{16}$$

$$\gamma_{A,AB} = E[D_A(11)] - E[D_A(10)] = P(c, \cdot) - P(s, \cdot) = P(j, \cdot). \tag{17}$$

Similar arguments yield

$$\gamma_{B0} = 0, \quad \gamma_{BA} = 0, \quad \gamma_{BB} = P(\cdot, s), \quad \gamma_{B,AB} = P(\cdot, j) \quad \text{and}$$
$$\gamma_{AB,0} = 0, \quad \gamma_{AB,A} = 0, \quad \gamma_{AB,B} = 0, \quad \gamma_{B,AB} = P(c, c).$$

Thus, the predicted values from the first stage are simply $\hat{D}_A = \gamma_A' Z$, $\hat{D}_B = \gamma_B' Z$ and $\widehat{D_A D_B} = \gamma_{AB}' Z$.

**Second stage** The second stage of the IV estimator consists of further regressing $Y$ on the linear projections $\hat{D}_A$, $\hat{D}_B$ and $\widehat{D_A D_B}$. More specifically, write

$$\begin{aligned} Y &= \beta_0 + \beta_A \hat{D}_A + \beta_B \hat{D}_B + \beta_{AB} \widehat{D_A D_B} + U_Y \\ &= \beta_0 + \beta_A \gamma_A' Z + \beta_B \gamma_B' Z + \beta_{AB} \gamma_{AB}' Z + U_Y \\ &= \beta_0 + (\beta_A, \beta_B, \beta_{AB}) \begin{pmatrix} \gamma_A' Z \\ \gamma_B' Z \\ \gamma_{AB}' Z \end{pmatrix} + U_Y \\ &= \beta_0 + \beta' \Gamma' Z + U_Y, \end{aligned} \tag{18}$$

where $\beta = (\beta_A, \beta_B, \beta_{AB})'$, $\Gamma = (\gamma_A, \gamma_B, \gamma_{AB})$ is the $3 \times 3$ matrix

$$\Gamma = \begin{pmatrix} P(s,\cdot) & 0 & 0 \\ 0 & P(\cdot,s) & 0 \\ P(j,\cdot) & P(\cdot,j) & P(c,c) \end{pmatrix},$$

and $E(\hat{D}_A Z) = E(\hat{D}_B Z) = E(\widehat{D_A D_B} Z) = 0$. As $\gamma_A$, $\gamma_B$, $\gamma_{AB}$ are linearly independent under Assumption ***, these orthogonality conditions are equivalent to $E[ZU_Y] = 0$. This means that equation (18) is also the linear projection of $Y$ on $Z$ and a constant, i.e., it can be identified with the reduced form regression

$$Y = \pi_0 + \pi_A Z_A + \pi_B Z_B + \pi_{AB} Z_A Z_B + U_Y. \tag{19}$$

**Comparison of the reduced form with the second stage** Theorem *** follows from comparing equation (18) with equation (19) and using the interpretation of the reduced form regression coefficients as intention to treat effects. In particular, (18) and (19) imply

$$\beta' \Gamma' = (\pi_A, \pi_B, \pi_{AB}) \Leftrightarrow \beta = \Gamma^{-1} \begin{pmatrix} \pi_A \\ \pi_B \\ \pi_{AB} \end{pmatrix}, \tag{20}$$

where $\Gamma^{-1}$ is given by

$$\Gamma^{-1} = \begin{pmatrix} \frac{1}{P(s,\cdot)} & 0 & 0 \\ 0 & \frac{1}{P(s,\cdot)} & 0 \\ -\frac{P(j,\cdot)}{P(s,\cdot)P(c,c)} & -\frac{P(\cdot,j)}{P(\cdot,s)P(c,c)} & \frac{1}{P(c,c)} \end{pmatrix}.$$

The second equation under (20) then yields

$$\beta_A = \frac{\pi_A}{P(s,\cdot)},\ \beta_B = \frac{\pi_B}{P(\cdot,s)},\ \beta_{AB} = \frac{1}{P(c,c)} \left[ \pi_{AB} - \pi_A \frac{P(j,\cdot)}{P(s,\cdot)} - \pi_B \frac{P(\cdot,j)}{P(\cdot,s)} \right]. \tag{21}$$

The reduced form coefficients $\pi_A$, $\pi_B$, $\pi_{AB}$ are given by formulas analogous to equations (11), (12) and (13):

$$\begin{aligned}
\pi_A &= E(Y|\, Z_A = 1, Z_B = 0) - E(Y|\, Z_A = 0, Z_B = 0) && (22)\\
\pi_B &= E(Y|\, Z_A = 0, Z_B = 1) - E(Y|\, Z_A = 0, Z_B = 0) && (23)\\
\pi_{AB} &= E(Y|\, Z_A = 1, Z_B = 1) - E(Y|\, Z_A = 0, Z_B = 1)\\
&- \big[E(Y|\, Z_A = 1, Z_B = 0) - E(Y|\, Z_A = 0, Z_B = 0)\big]. && (24)
\end{aligned}$$

Thus, $\beta_A = \pi_A / P(s, \cdot)$ is precisely the Wald estimand (1), and $\beta_B = \pi_B / P(\cdot, s)$ is the analogous Wald estimand for treatment $B$. The causal interpretation of these quantities has already been given in Theorem 1 and is restated in Theorem 3.

Deriving the interpretation of $\pi_{AB}$ is more complicated. This coefficient is the difference between two intention-to-treat effects: $ITT_{A|B}$, given by the numerator in (2), and $ITT_{A|\bar{B}}$, given by the numerator in (1). Hence, by Theorems 1 and 2,

$$\begin{aligned}
\pi_{AB} &= ATE_{AB}(c,j)P(c,j) + ATE_{A|B}(c,s)P(c,s) + ATE_{A|\bar{B}}(c,n)P(c,n)\\
&\quad -ATE_{A|\bar{B}}(s,\cdot)P(s,\cdot). && (25)
\end{aligned}$$

By Assumption 3, $P(c,n) = P(s,n)$ and $ATE_{A|\bar{B}}(c,n) = ATE_{A|\bar{B}}(s,n)$. Furthermore,

$$ATE_{A|\bar{B}}(s,\cdot)P(s,\cdot) = ATE_{A|\bar{B}}(s,c)P(s,c) + ATE_{A|\bar{B}}(s,n)P(s,n).$$

Substituting the preceding expressions into (25) gives

$$\pi_{AB} = ATE_{AB}(c,j)P(c,j) + ATE_{A|B}(c,s)P(c,s) - ATE_{A|\bar{B}}(s,c)P(s,c). \quad (26)$$

We can decompose the first term in (26) as

$$\begin{aligned}
ATE_{AB}(c,j)P(c,j) &= E\big\{[Y(11) - Y(01) + Y(01) - Y(00)]\\
&\quad \times 1(D_A(11) = 1, D_B(01) = 0, D_B(11) = 1)\big\}\\
&= ATE_{A|B}(c,j)P(c,j) + ATE_{B|\bar{A}}(c,j)P(c,j). && (27)
\end{aligned}$$

Substituting (27) into (26) yields

$$\begin{aligned}
\pi_{AB} &= ATE_{A|B}(c,c)P(c,c) + ATE_{B|\bar{A}}(c,j)P(c,j) - ATE_{A|\bar{B}}(s,c)P(s,c)\\
&= [ATE_{A|B}(c,c) - ATE_{A|\bar{B}}(c,c)]P(c,c)\\
&\quad +ATE_{A|\bar{B}}(j,c)P(j,c) + ATE_{B|\bar{A}}(c,j)P(c,j), && (28)
\end{aligned}$$

where the first equality uses the fact

$$ATE_{A|B}(c,j)P(c,j) + ATE_{A|B}(c,s)P(c,s) = ATE_{A|B}(c,c)P(c,c)$$

and the second equality uses the fact

$$ATE_{A|\bar{B}}(c,j)P(c,j) + ATE_{A|\bar{B}}(c,s)P(c,s) = ATE_{A|\bar{B}}(c,c)P(c,c).$$

Substituting (28) into (21) then gives

$$\begin{aligned}
\beta_{AB} &= \frac{1}{P(c,c)}\Big\{[ATE_{A|B}(c,c) - ATE_{A|\bar{B}}(c,c)]P(c,c) + ATE_{A|\bar{B}}(j,c)P(j,c) \\
&\quad + ATE_{B|\bar{A}}(c,j)P(c,j) - ATE_{A|\bar{B}}(s,\cdot)P(j,\cdot) - ATE_{B|\bar{A}}(\cdot,s)P(\cdot,j)\Big\} \\
&= ATE_{A|B}(c,c) - ATE_{A|\bar{B}}(c,c) + [ATE_{A|\bar{B}}(j,\cdot) - ATE_{A|\bar{B}}(s,\cdot)]\frac{P(j,\cdot)}{P(c,c)} \\
&\quad + [ATE_{B|\bar{A}}(\cdot,j) - ATE_{B|\bar{A}}(\cdot,s)]\frac{P(\cdot,j)}{P(c,c)},
\end{aligned}$$

where we use the fact that $P(j,\cdot) = P(j,c)$ and $P(\cdot,j) = P(c,j)$ by Assumption 3. This is the expression for $\beta_{AB}$ given in Theorem 3.

## References


Angrist, J.D. and Pischke, J-S. (2009). *Mostly Harmless Econometrics.* Princeton University Press, Princeton, New Jersey.

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.

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

Donald, S. G., Y.-C. Hsu and R. P. Lieli (2014). "Testing the Unconfoundedness Assumption via Inverse Probability Weighted Estimators of (L)ATT," *Journal of Business and Economic Statistics*, 32, 395-415.

Hudgens, M.G. and Halloran, M.E. (2008). "Toward Causal Inference with Interference," *Journal of the American Statistical Association*, 103, 832-842.

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

Kang, H. and Imbens, G. (2016). "Peer Encouragement Designs in Causal Inference with Partial Interference and Identification of Local Average Network Effects." Working paper.

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," *The 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.