Instrumental Variable Methods for Conditional Effects and Causal Interaction in Voter Mobilization Experiments[*]

Matthew Blackwell[†]

Abstract

In democratic countries, voting is one of the most important ways for citizens to influence policy and hold their representatives accountable. And yet, in the United States and many other countries, rates of voter turnout are alarmingly low. Every election cycle, mobilization efforts encourage citizens to vote and ensure that elections reflect the true will of the people. To establish the most effective way of encouraging voter turnout, this paper seeks to differentiate between (1) the *synergy hypothesis* that multiple instances of voter contact increase the effectiveness of a single form of contact, and (2) the *diminishing returns hypothesis* that multiple instances of contact are less effective or even counterproductive. Remarkably, previous studies have been unable to compare these hypotheses because extant approaches to analyzing experiments with noncompliance cannot speak to questions of causal interaction. I resolve this impasse by extending the traditional instrumental variables framework to accommodate multiple treatment-instrument pairs, which allows for the estimation of conditional and interaction effects to adjudicate between synergy and diminishing returns. The analysis of two voter mobilization field experiments provides the first evidence of diminishing returns to follow-up contact and a cautionary tale about experimental design for these quantities.

[*] Thanks to Peter Aronow, Fred Boehmke, Jake Bowers, Alex Coppock, Don Green, Erin Hartman, Kosuke Imai, Luke Keele, Gary King, Michael LaCour, Jamie Monogan, James Robins, Arturas Rozenas, Maya Sen, Susumu Shikano, Tyler VanderWeele, and Teppei Yamamoto for helpful suggestions and feedback. Special thanks to David Nickerson for sharing his data. All remaining errors are my own.

[†] Assistant Professor, Department of Government, Harvard University, Institute for Quantitative Social Science, 1737 Cambridge Street, Cambridge, MA 02138 (email: `mailto:mblackwell@gov.harvard.edumblackwell@gov.harvard.edu`, web: `mattblackwell.org`)

1 Introduction

In democratic countries, casting a ballot is one of the most important ways for ordinary citizens to influence policy or hold their representatives accountable. Low voter turnout can skew election results away from the true “will of the people” and result in policies that disproportionately benefit a minority of citizens (see Fowler, 2013, and references therein). To combat or take advantage of this, in every election cycle nonprofit groups, political parties, and candidates themselves dedicate tremendous resources to get-out-the-vote (GOTV) efforts. Over the course of the 2012 presidential election, the Barack Obama campaign alone had 2.2 million volunteers make over 150 million door knocks and phone calls designed to encourage (supportive) citizens to vote (Obama for America, 2013), which is remarkable since there were only 127 million votes actually cast in the election.

To combat the effects of low voter turnout on the electoral process, it is crucial to understand the efficacy of these types of GOTV contact and how they might interact. For example, does door-knocking increase turnout? Should it be paired with phone calls or done in isolation? Two rival hypotheses exist in the GOTV literature to answer this latter question: (1) the *synergy hypothesis* that multiple forms of contact combine to produce larger effects than either separately, and (2) the *diminishing returns hypothesis* that additional contact is less effective or even counterproductive compared to a single intervention (Green, McGrath and Aronow, 2013). These two hypotheses have strong implications for how to maximize voter turnout with a fixed budget. If diminishing returns exist, then GOTV campaigns should focus their efforts on maximizing the number of citizens contacted, whereas if synergy exists, they should attempt more follow-up contact on a smaller group of respondents.

The synergy and diminishing returns hypotheses directly concern the *causal interaction* between multiple forms of contact, a quantity well-understood in traditional randomized experiments with full compliance (VanderWeele, 2015). Unfortunately, experimental studies of GOTV efforts are subject to noncompliance since citizens often fail to answer their door or pick up the phone, which could lead to unmeasured confounding between actual (as opposed to randomized) GOTV contact and voter turnout. Thus, while a large and influential literature in political science has relied on randomized experiments to assess the impact GOTV appeals (Gerber and Green, 2000; Green, Gerber and Nickerson, 2003; Nickerson, 2007; Gerber, Green and Larimer, 2008; Arceneaux and Nickerson, 2009), there has been no study of the interaction between different forms of GOTV contact that has dealt with issues of noncompliance. Gerber and Green (2000) use an intention-to-treat (ITT) analysis to investigate the interaction between the randomization to in-person canvassing and the randomization to telephone canvassing, but noncompliance makes it difficult to ascertain if these ITT interactions reflect the true interaction between *actual* in-person contact and *actual* telephone contact. Furthermore, there is no statistical literature on how one would use instrumental variables, a common approach to handling noncompliance, to estimate these causal interactions.

This paper extends the single-treatment instrumental variables framework of Angrist, Imbens and Rubin (1996) to allow for the identification and estimation of certain conditional and interaction effects of two forms of GOTV contact. Each binary randomized assignment to GOTV contact serves as an instrument for the binary treatment of actual GOTV contact and the goal is to identify the joint effects and interactions of the binary treatments. Previous extensions of IV methods to allow for multiple-level treatments (relevant since two binary treatments can be thought of as a single four-level treatment) did not separately identify conditional effects, but rather a weighted average that would be difficult to interpret in this context (Angrist and Imbens, 1995). In contrast, I use a principal stratification approach to generalize the concept of compliance types to the setting of two binary treatments and define various "local" quantities of interest, which are average comparisons between pairs of treatment vectors (that is, actual contact) within levels of these compliance types. These include a local average interaction effect (LAIE) that I use to adjudicate between the synergy and diminishing returns hypotheses. I also show that these estimands are nonparametrically identified from the usual IV assumptions and a novel assumptions called the *treatment exclusion restriction*, which ensures that each treatment is only affected by a single instrument. The paper develops consistent and asymptotically normal plug-in estimators for these quantities along with consistent variance estimators. I also show how an interacted version of two-stage least squares (TSLS) can identify the LAIE but cannot identify conditional effects without further assumptions.

I investigate the question of synergy versus diminishing returns in two large GOTV field experiments that can speak to these questions. In the first, over 11,000 households were randomly assigned to receive telephone canvassing, in-person canvassing, both, or neither ahead of the 1998 general election in New Haven, CT (Gerber and Green, 2000). In this experiment, there is strong variation in conditional effects between compliance groups, indicating that compliance with phone contact, for instance, may be predictive of response to door-to-door canvassing. But this experiment also suffered from very low rates of compliance and an experimental design that minimized the number of respondents that received both forms of treatment. These factors combined to produce very high uncertainty over the conditional and interaction effects.

To overcome issues of low compliance, I analyze a second experiment, conducted by the Youth Voting Coalition (YVC) during the 2004 presidential election, wherein over 26,000 young registered voters across 10 sites were randomly assigned to phone calls from either a volunteer phone bank, a professional phone bank, both, or neither (Nickerson, 2007). The compliance rate in this experiment was far higher than in the New Haven experiment. I find support for the diminishing returns hypothesis among college-aged students—a second round of phone contact produces far weaker mobilizing effects than a single phone call. This represents the first piece of evidence for causal interaction in GOTV efforts. Finally, I discuss how the estimands developed in this paper can be useful for maximizing the cost effectiveness of future GOTV efforts.

2 A Joint Instrumental Variable Framework for GOTV Studies

Using the New Haven experiment as a guide, let $D_{i1} = 1$ indicate that a respondent received a phone canvassing message and $D_{i2} = 1$ indicate that a respondent received in-person canvassing. Furthermore, let $Z_{i1} \in \{0, 1\}$ be random assignment to phone canvassing and Z_{i2} be random assignment to in-person canvassing. Let Y_i be the outcome of interest, whether or not respondent i voted. Define $Y_i(d_1, d_2)$ as the potential outcomes, representing whether or not respondent i would have voted if we set D_{i1} to d_1 and D_{i2} to d_2 (Rubin, 1974). Note that under the usual exclusion restriction (stated formally below), these potential outcomes will not depend on the value of the instruments.

In these experiments, canvassers were unable to contact all respondents assigned to receive contact. For phone contact, this could be caused by incorrect phone numbers in voter registration records or by respondents refusing to answer the phone. For in-person canvassing, the travel time between physical households could be quite long and could prevent the canvassers from reaching all households intended for contact (Gerber and Green, 2000, p. 655). Furthermore, canvassing in the New Haven experiment ceased at sunset for safety reasons, preventing the canvassers from contacting residents not home during the day. So in these studies D_{ij} may not equal Z_{ij}. To account for this, define the potential (outcome for) treatment by $D_{i1}(z_1, z_2)$ as the outcome which would be observed if the experimenter set $(Z_{i1}, Z_{i2}) = (z_1, z_2)$, with $D_{i2}(z_1, z_2)$ similarly defined. By using this definition, I have implicitly assumed the potential treatments with respect to one treatment are unaffected by the value taken by the other, so that $D_{i1}(z_1, z_2, d_1) = D_{i1}(z_1, z_2, d_1') = D_{i1}(z_1, z_2)$, and similarly for D_{i2}. It is also useful to define $D_{i1}(z_1) = D_{i1}(z_1, Z_{i2})$ and $D_{i2}(z_2) = D_{i2}(Z_{i1}, z_2)$ as the potential outcomes for these variables when only setting one of the two instruments and leaving the other instrument at its observed value. Finally, let $D_i = (D_{i1}, D_{i2})$ and $Z_i = (Z_{i1}, Z_{i2})$. Implicit in the definition of all these potential outcomes is the stable unit treatment value assumption (SUTVA) of no interference between units since the potential outcomes and treatments for unit i only depend on the treatment and instrument values of unit i. Note that in these types of GOTV experiments, there is often, though not always, one-sided noncompliance whereby if $Z_i = 0$ then $D_i = 0$ since treatment is withheld from those who are assigned to control. I develop the statistical theory below without relying on this assumption and highlight where it simplifies the general framework.

This paper generalizes the single-treatment IV framework of Angrist, Imbens and Rubin (1996), which makes the following assumptions modified for the presence of two binary treatments:

Assumption 1 (Single-Treatment IV). *For all values $d_1, d_2, z_1, z_1', z_2, z_2'$ and $j \in \{1, 2\}$, the following statements hold:*

1. *Consistency:* $Y_i = Y_i(D_{i1}, D_{i2})$, $D_{i1} = D_{i1}(Z_{i1}, Z_{i2})$, *and* $D_{i2} = D_{i2}(Z_{i1}, Z_{i2})$*;*

2. *Randomization of the instruments:* $(Z_{i1}, Z_{i2}) \perp\!\!\!\perp \left\{Y_i(d_1, d_2), D_{i1}(z_1, z_2), D_{i2}(z'_1, z'_2)\right\}$.

3. *Outcome exclusion restriction:* $Y_i(d_1, d_2, z_1, z_2) = Y_i(d_1, d_2, z'_1, z'_2) = Y_i(d_1, d_2)$.

4. *Effect of instrument on treatment:* $\Pr[D_{ij}(z_j = 1)] \neq \Pr[D_{ij}(z_j = 0)]$.

5. *Monotonicity:* $\Pr[D_{ij}(z_j = 1) \geq D_{ij}(z_j = 0)] = 1$.

These are the standard assumptions of Angrist, Imbens and Rubin (1996) applied separately to each instrument-treatment pair and allowing for the joint randomization of Z_i. The exclusion restriction is uncontroversial in these GOTV studies since respondents only learn of their random assignment through actual contact. Monotonicity is often, though not always, satisfied by restricting contact to those who are randomly assigned to receive it, which ensures one-sided noncompliance. For the sake of generality, I develop the framework using the more general monotonicity assumption. Note that this framework depends heavily on binary instruments and treatment.

Under Assumption 1, there may be unmeasured confounding between actual (as opposed to assigned) GOTV contact and voter turnout, which renders the average treatment effects of actual contact (D_{ij}) unidentified. Angrist, Imbens and Rubin (1996) instead identified average causal effects within levels of a respondent's *compliance type* or *principal strata*, which are defined by how a unit responds to the instrument (Frangakis and Rubin, 2002). In the single-treatment IV case with a binary treatment and instrument, these strata are defined by the vector $(D_i(0), D_i(1))$, which can respectively take on four different values: compliers with $(0, 1)$, always-takers with $(1, 1)$, never-takers with $(0, 0)$, and defiers with $(1, 0)$. Using this notation, let $C_i \in \{c, a, n, d\}$ denote the compliance type for unit i. Monotonicity implies that there are no defiers in the population and allows for the identification of the local average treatment effect (LATE), which is the average effect of actual contact among compliers: $E[Y_i(1) - Y_i(0)|C_i = c]$. Past analyses of GOTV experiments have leveraged this fact to estimate the LATEs of each form of actual GOTV contact separately.

In the GOTV context, compliers are those respondents that are *reachable*—they pick up the phone when called or answer the door when canvassers knock.

While this basic framework has been previously extended beyond single treatments, these endeavors have not addressed the identification or estimation of causal interactions or conditional effects in 2×2 factorial experiments such as the ones considered here. Angrist and Imbens (1995) extended the single-treatment setting to accommodate multiple-level instruments and treatments, focusing on ordinal treatments, where higher values indicate "more" treatment. While it is possible to redefine two binary treatments as a single four-level treatment, their results cannot identify the separate effects of between-level contrasts ($Y_i(1,0)$ vs. $Y_i(0,0)$, for example), but rather a complicated weighted average of these comparisons. Cheng and Small (2006) study noncompliance in randomized experiments to compare the effectiveness of two treatments, but they focus on three-arm trials where both treatments are never assigned together not the full factorial designs of these GOTV studies. Finally, an alternative approach to IV uses a class of models called structural nested mean models (SNMMs), avoiding monotonicity and instead relying homogeneity assumptions on the causal effects to point identify population-average effects (for example, Robins, 1994; Hernán and Robins, 2006; Clarke and Windmeijer, 2010). In the present context, monotonicity holds by design and so I focus on an approach that leverages this fact.

2.1 Joint compliance types, treatment exclusion, and quantities of interest

To generalize single-treatment IV framework, I will define, identify, and estimate average causal effects of actual contact within levels of the two-treatment principal strata. The definition of these principal strata with two forms of contact, which I call the *joint compliance types*, follows a similar logic to the single-treatment case and can be written as a vector of 8 values:

$$(D_{i1}(0,0), D_{i1}(1,0), D_{i1}(0,1), D_{i1}(1,1), D_{i2}(0,0), D_{i2}(1,0), D_{i2}(0,1), D_{i2}(1,1))\,.$$

This vector completely characterizes how both potential treatments react to both instruments. The joint compliance types can be reorganized in terms of the single-treatment compliance types by noting that, for example, $\{D_{i1}(0,0), D_{i1}(1,0)\}$ defines the compliance type for phone contact when

i is assigned to no in-person contact. This vector can take on the four above types. For example, if $D_{i1}(0,0) = 0, D_{i2}(1,0) = 1$ then i is a complier for D_{i1} when $z_2 = 0$. More generally, let $C_{i1|z} \in \{c, a, n, d\}$ be the compliance status for D_{i1} when $z_2 = z$ (that is, for $D_{i1}(0,z), D_{i1}(1,z)$) and $C_{i2|z}$ be the compliance status for D_{i2} when $z_1 = z$ (that is, for $D_{i2}(z,0), D_{i2}(z,1)$). We can then characterize the joint compliance types for unit i as $C_i = (C_{i1|0}, C_{i1|1}, C_{i2|0}, C_{i2|1})$, letting $C_i = kmst$ be a shorthand for $C_i = (k, m, s, t)$ where $k, m, s, t \in \{c, a, n, d\}$. With four elements taking on four different values, there are $4^4 = 256$ of these principal strata. One example of these compliance types is the "joint complier" group, $C_i = cccc$, who are reachable by either phone or door-to-door canvassing under any randomization. Another group are those with $C_i = ncnc$ that we might call the "interacted compliers" and who can only be contacted when they are assigned to receive both forms of contact: $D_i(z_1, z_2) = (1,1)$ if $(z_1, z_2) = (1,1)$ and $(0,0)$ otherwise. For the joint compliers, each treatment only depends on its own instrument, whereas for the interacted compliers, D_{i1} depends on z_2 and D_{i2} depends on z_1.

To identify and estimate the causal effect of multiple types of actual GOTV contact within joint compliance types, two sets of causal quantities need to be identified: the average potential outcomes within principal strata, $E[Y_i(d_1, d_2)|C_i = kmst]$, and the probability of each principal strata, $\Pr[C_i = kmst]$. To identify these, we will use the observed-data parameters, which consist of the 12 treatment-instrument strata probabilities, $\Pr[D_i = (d_1, d_2)|Z_i = (z_1, z_2)]$, and the 16 average observed outcomes, $E[Y_i|D_i = (d_1, d_2), Z_i = (z_1, z_2)]$. These 28 observed-data parameters will clearly not be sufficient to identify the 255 compliance probabilities, let alone the full set of causal parameters. Monotonicity in the present setting reduces the number of compliance types to $3^4 = 81$, since there are no defiers, but even this is not enough to achieve identification.

To reduce the number of compliance types even further and identify the causal parameters, I rely on a novel assumption that is justified from the experimental design of the GOTV studies. Namely, if the two forms of canvassing are done by different organizations (as is true both experiments) and if the forms of contact are substantially different (as they are in the New Haven experiment), then it is reasonable to assume that randomization to one form of contact has no impact on the actual receipt of the other form of contact. I call this assumption the *treatment exclusion restriction*:

Assumption 2 (Treatment Exclusion Restriction)**.** *For all values* z_1, z_1', z_2, z_2',

$$D_{i1}(z_1, z_2) = D_{i1}(z_1, z_2') = D_{i1}(z_1),$$
$$D_{i2}(z_1, z_2) = D_{i2}(z_1', z_2) = D_{i2}(z_2).$$

Treatment exclusion says that each instrument only affects its "own" treatment and so the compliance type for one form of contact is unaffected by the randomization of the other contact. This implies that $C_{ij|0} = C_{ij|1} = C_{ij}$ for $j \in \{0,1\}$ and allows us to simplify the principal strata as $C_i = km$ (meaning $C_{i1} = k$ and $C_{i2} = m$), leaving $3^2 = 9$ compliance types and making identification possible. Table 1 shows the mapping between values C_i and the potential contacts, $D_{ij}(z_j)$. In the GOTV experiments, treatment exclusion could be violated if, for instance, actually receiving a phone call changed a person's propensity to respond to attempts at door-to-door contact. This could occur if a respondent finds a single conversation with an in-person canvasser uncomfortable, discouraging them from answering the phone in future contact attempts so that $C_{i1|0} = c$ and $C_{i1|1} = n$. While this is a strong assumption and must be evaluated in each context, one silver lining is that, under Assumptions 1(i) and (ii), it has testable implications. Conditional on its own instrument, a treatment should be unrelated to the other instrument so that $E[D_{i1}|Z_{i1} = z_1, Z_{i2} = 1] = E[D_{i1}|Z_{i1} = z_1, Z_{i2} = 0]$ and $E[D_{i2}|Z_{i1} = 1, Z_{i2} = z_2] = E[D_{i2}|Z_{i1} = 0, Z_{i2} = z_2]$, which can be tested with, say, a regression of each treatment on both instruments to detect violation of the treatment exclusion restriction.

We can now define the quantities of interest in these GOTV experiments. Their factorial designs allow the investigation of the conditional effect of one form of contact conditional on the other, but noncompliance only allows for identification of these quantities within the above principal strata. I call these the *local average conditional effects*, or LACEs. Each LACE is associated with a particular treatment, a particular level of the other treatment, and a joint compliance type:

$$\tau_{km,1}(d_2) = E[Y_i(1, d_2) - Y_i(0, d_2)|C_i = km],$$

$$\tau_{km,2}(d_1) = E[Y_i(d_1, 1) - Y_i(d_1, 0)|C_i = km],$$

recalling that $C_i = km$ corresponds to $(C_{i1}, C_{i2}) = (k, m) \in \{c, n, a\}$. This is the contrast between receiving treatment j and not receiving it with the other treatment fixed at a particular value among those in stratum C_i. There are four basic conditional effects with two binary treatments, but not all of them are well-defined for all compliance types. For instance, the contrast $Y_i(1, 0) - Y_i(0, 0)$ is meaningless for always-takers for the second treatment, because this group would logically never observe $Y_i(1, 0)$ or $Y_i(0, 0)$ since they will never have $D_{i2} = 0$. In general, if $C_{i1} = n$ then $Y(1, .)$ can never be observed, and if $C_{i1} = a$ then $Y(0, .)$ can never be observed, and similarly for C_{i2}. Table 1 shows the basic LACEs in this setup—four for the joint compliers, and one for each of the strata that comply with at least one of the treatments.

The LACEs can provide useful information both to political scientists about how voters make decisions and to candidates and campaigns who can use them to better allocate their resources. For instance, if a campaign has made a failed attempt to call a citizen, then they know that because of one-sided noncompliance this person is a never-taker for phone calls and can base their decision to canvass their physical address on the LACE that conditions on this information, $\tau_{nc,2}(0)$. The campaign obviously will not know if this respondent is a complier for in-person canvassing, but can use the compliance type probabilities from past studies to derive a prediction for the cost-effectiveness of a second GOTV effort.

The LACEs can be combined to create other interesting estimands. The synergy and diminishing returns hypotheses, for instance, speak to the causal interaction between the two treatments. Thus, I define the *local average interaction effect*, or LAIE, which is the difference in the conditional effects for the joint compliers:

$$\tau_{LAIE} = \tau_{cc,1}(1) - \tau_{cc,1}(0) = E[(Y_i(1, 1) - Y_i(0, 1)) - (Y_i(1, 0) - Y_i(0, 0))|C_i = cc]. \quad (1)$$

This represents the difference in the effect of phone canvassing when it is and is not paired with in-person canvassing, which is the exact quantity needed to adjudicate these two hypotheses. Synergy would imply that $\tau_{LAIE} > 0$, while diminishing returns would imply that $\tau_{LAIE} < 0$. In addition to the LAIE, I also define the *local average joint effect*, or LAJE, as the effect of receiving both treatments for the joint compliers: $\tau_{LAJE} = \tau_{cc,1}(1) + \tau_{cc,2}(0) = E[Y_i(1,1) - Y_i(0,0)|C_i = cc]$. Both of these quantities are only defined among the joint compliers since they are functions of multiple LACEs. If a respondent is not reachable by both forms of contact, it is hardly sensible to discuss the interaction or joint effects for them.

It is also possible to aggregate LACEs across compliance types, which can be useful to increase power and to maximize the target population of interest. To do this, we simply take the weighted average of the basic LACEs across the compliance types for which they are defined:

$$\tau_{c-,1}(0) = E[Y_i(1,0) - Y_i(0,0)|C_i \in \{cc, cn\}] = (\rho_{cc} + \rho_{cn})^{-1} \left(\rho_{cc}\tau_{cc,1}(0) + \rho_{cn}\tau_{cn,1}(0)\right). \qquad (2)$$

As we will see below, these weighted averages are the target of TSLS estimation in designs with one-sided noncompliance.

3 Identification and Estimation

To identify these estimands, we must connect the causal parameters with the parameters of the observed data. Table 2 provides a sense of how this works by showing how the joint compliance type probabilities, $\rho_{km} = \Pr[C_i = km]$, relate to the probability of strata defined by combinations of instruments and treatments, $f_{d_1 d_2|z_1 z_2} = \Pr[D_i = (d_1, d_2)|Z_i = (z_1, z_2)]$, under Assumptions 1 and 2. For instance, one strata of subjects were never contacted and so have $D_i = (0,0)$, but were randomized to receive both forms of contact and so have $Z_i = (1,1)$. These respondents are uniquely identified as never-takers on both treatments:

$$\begin{aligned} f_{00|11} &= \Pr[D_i = (0,0)|Z_i = (1,1)] = \Pr[D_{i1}(1) = 0, D_{i2}(1) = 0|Z_i = (1,1)] && (3) \\ &= \Pr[D_{i1}(1) = 0, D_{i2}(1) = 0] = \rho_{nn} && (4) \end{aligned}$$

The first equality comes from consistency, the second from randomization, and the last can be established by noting that *nn* is the only compliance type in Table 1 that has $D_{ij}(1) = 0$ for all j. Other observed strata are a mixture over several compliance types. We can use similar ideas to connect observed-outcome means to the means of the potential outcomes within principal strata, $E[Y_i(d_1, d_2)|C_i = km]$. Theorem 1 combines these ideas to identify the basic LACEs.

Theorem 1. *Under Assumptions 1 and 2, all of the basic LACEs in Table 1 are nonparametrically identified as:*

$$\begin{array}{ll}
\tau_{cc,1}(0) = \dfrac{S^1_{0|10} - S^1_{0|11} - S^1_{0|00} + S^1_{0|01}}{f_{11|11} - f_{11|01} - f_{11|10} + f_{11|00}} & \tau_{cc,1}(1) = \dfrac{S^1_{1|11} - S^1_{1|10} - S^1_{1|01} + S^1_{1|00}}{f_{11|11} - f_{11|01} - f_{11|10} + f_{11|00}} \\
\tau_{cn,1}(0) = \dfrac{S^1_{0|11} - S^1_{0|01}}{f_{10|11} - f_{10|01}} & \tau_{ca,1}(1) = \dfrac{S^1_{1|10} - S^1_{1|00}}{f_{11|10} - f_{11|00}} \\
\tau_{cc,2}(0) = \dfrac{S^2_{0|01} - S^2_{0|11} - S^2_{0|00} + S^2_{0|10}}{f_{11|11} - f_{11|01} - f_{11|10} + f_{11|00}} & \tau_{cc,2}(1) = \dfrac{S^2_{1|11} - S^2_{1|10} - S^2_{1|01} + S^2_{1|00}}{f_{11|11} - f_{11|01} - f_{11|10} + f_{11|00}} \\
\tau_{nc,2}(0) = \dfrac{S^2_{0|11} - S^2_{0|10}}{f_{01|11} - f_{01|10}} & \tau_{ac,2}(1) = \dfrac{S^2_{1|01} - S^2_{1|00}}{f_{11|01} - f_{11|00}}
\end{array}$$

where $S^j_{d|z_1 z_2} = E\left[Y_i \mathbb{I}\left(D_{i,-j} = d\right)|Z_i = (z_1, z_2)\right]$, *and* $\mathbb{I}(\cdot)$ *is an indicator function.*

The proof for Theorem 1 is given in the Supplemental Materials, which follows a similar strategy to Abadie (2003). The quantities in the numerator for the LACEs of treatment j, $S^j_{d|z_1 z_2}$, are expectations of the outcome multiplied by an indicator *for the other instrument*, which helps select out the appropriate compliance types. Thus, the numerators in these results represent ITT effects of the phone-call and door-to-door randomizations on transformations of the outcome. The denominators represent the estimated probability of being in a particular joint compliance stratum. Theorem 1 implies that the LAIE, LAJE and aggregated LACEs are also identified, since they are simple combinations of the LACEs. The LAIE in particular has an interpretable form. Letting $S_{z_1 z_2} = E[Y|Z_i = (z_1, z_2)]$, we have:

$$\tau_{LAIE} = \frac{(S_{11} - S_{01}) - (S_{01} - S_{00})}{(f_{11|11} - f_{11|01}) - (f_{11|10} - f_{11|00})} \qquad (5)$$

Thus, the LAIE is the ratio of the ITT interaction for Z_i and Y_i and an ITT interaction between Z_i

and D_i.

All of these identification results have a similar structure: a function of reduced-form relationships divided by a probability of various compliance strata. One potential concern for all of these results is that of weak instruments, where the relationship between the instrument and the treatment is weak (Bound, Jaeger and Baker, 1995). In the case of a binary instrument and a binary treatment, this is equivalent to low probability of compliance. Thus, estimates of the basic LACEs might have higher variance than the LATE for either of the two treatments separately. This is one reason to focus on the aggregated LACEs compared to ones that focus only a single compliance class. This problem is especially salient in the New Haven experiment below.

Theorem 1 not only identifies the LACEs, the LAJE, and the LAIE, but it also suggests an estimation strategy. With a random sample of N units, $(Y_i, D_{i1}, D_{i2}, Z_{i1}, Z_{i2})$, from an infinite super-population, it is possible to estimate each of the expectations above. For example, we replace the quantity $S^1_{d|z_1z_2}$, with its sample analogue,

$$\widehat{S}^1_{d|z_1z_2} = \frac{\sum_{i=1}^N Y_i \mathbb{I}(D_{i2} = d)\mathbb{I}(Z_{i1} = z_1)\mathbb{I}(Z_{i2} = z_2)}{\sum_{i=1}^N \mathbb{I}(Z_{i1} = z_1)\mathbb{I}(Z_{i2} = z_2)}.$$

If we replace all expectations in Theorem 1 with their sample quantities, we obtain plug-in estimators for all of the quantities of interest. In the Supplemental Materials, I show that, under mild regularity conditions, these estimators are consistent and asymptotically normal and derive a closed-form expression for their asymptotic variance. If the randomization assumption only holds conditional on covariates, it is necessary to modify these estimators to include the covariates. For a set of discrete covariates, X_i, it is possible to calculate each of the above expectations conditional on levels of X_i and then average these stratum-specific effects over the distribution of X_i. I take this approach in the second experiment below, where randomization was stratified by site. With continuous covariates, one could replace the expectations with parametric models and average the estimates over the empirical distribution of X_i.

3.1 Two-stage least squares

In the original analyses of these GOTV experiments, the authors relied on two-stage least squares (TSLS) to estimate the separate effects of both forms of GOTV contact. This approach is justified by Imbens and Angrist (1994), Angrist and Imbens (1995), and Abadie (2003) who document the relationship between TSLS and the LATE. These studies have shown that the TSLS estimand *is* the LATE when the TSLS model includes no covariates and the instrument and treatment are binary. Unfortunately, these results are not immediately applicable to questions of causal interaction between phone and in-person contact. I extend TSLS to allow for interactions and investigate what local conditional or interactive effects it can estimate in the two-treatment setting.

Let **D** and **Z** be $N \times 4$ design matrices that contain a constant, each instrument or treatment, and their interaction. That is, let $\mathbf{D}_i = \begin{pmatrix} 1 & D_{i1} & D_{i2} & D_{i1}D_{i2} \end{pmatrix}$ be a generic row of **D**, and $\mathbf{Z}_i = \begin{pmatrix} 1 & Z_{i1} & Z_{i2} & Z_{i1}Z_{i2} \end{pmatrix}$ be a generic row of **Z** Let $\mathbf{Y}^\intercal = \begin{pmatrix} Y_1 & Y_2 & \cdots & Y_N \end{pmatrix}$ be the vector of outcomes. With these definitions in hand, it is possible to define the interacted TSLS estimator (iTSLS), which is simply the application of TSLS to the above design matrices: $\widehat{\delta} = (\mathbf{D}^\intercal\mathbf{Z})^{-1}\mathbf{Z}^\intercal\mathbf{Y}$. This estimator allows for an interaction between the two treatments and is fully saturated in both stages. Under standard regularity conditions, this estimator will converge to $\delta = \left(E[\mathbf{D}_i^\intercal\mathbf{Z}_i]\right)^{-1} E[\mathbf{Z}_\mathbf{i}{}^\intercal Y_i]$. Finally, let the limiting coefficient vector be labelled as $\delta^\intercal = \begin{pmatrix} \delta_0 & \delta_1 & \delta_2 & \delta_3 \end{pmatrix}$.

How does δ map onto the estimands defined above? The following theorem shows this connection in the case where the instruments are independent.

Theorem 2. *Suppose that Assumptions 1 and 2 hold and that $Z_{i1} \perp\!\!\!\perp Z_{i2}$ and let $\omega_{ck,1} = \rho_{ck}/(\rho_{cc} + \rho_{cn} + \rho_{ca})$ and $\omega_{kc,2} = \rho_{kc}/(\rho_{cc} + \rho_{nc} + \rho_{ac})$ for $k \in \{c, a, n\}$. Then,*

$$\delta_1 = \omega_{cc,1}\tau_{cc,1}(0) + \omega_{cn,1}\tau_{cn,1}(0) + \omega_{ca,1}\left[\tau_{ca,1}(1) - \tau_{LAIE}\right] \tag{6}$$

$$\delta_2 = \omega_{cc,2}\tau_{cc,2}(0) + \omega_{nc,2}\tau_{nc,2}(0) + \omega_{ac,2}\left[\tau_{ac,2}(1) - \tau_{LAIE}\right] \tag{7}$$

$$\delta_3 = \tau_{cc,1}(1) - \tau_{cc,1}(0) = \tau_{LAIE}. \tag{8}$$

In the Supplemental Materials, I provide a proof is provided and a simulation exercise that confirms the divergence between the plug-in estimators and the TSLS approach. The coefficient on the interaction between the two treatments is indeed the interaction effect for joint compliers, τ_{LAIE}. The lower order terms are more complicated. The coefficient on D_{i1}, for instance, is a mixture of three different effects, weighted by the probability of *cc*, *cn*, and *ca* (respectively) conditional being a complier for D_{i1}. The first and second components of this mixture are the LACEs of D_{i1} when $d_2 = 0$ for the joint compliers (*cc*) group and the complier-never-taker group (*cn*), respectively. But there is no such LACE for the *ca* group, so the third component of the mixture is the LACE of D_{i1} when $d_2 = 1$ for the *ca* group *minus the interaction effect for the joint compliers*. Essentially, TSLS is imputing $Y_i(1, 0) - Y_i(0, 0)$ for a group that could never see this effect (*ca*) by drawing information from the joint compliers about the interaction between the two treatments. This has two ramifications. First, the lower order terms will have no direct interpretation as a causal effect without further assumptions. Second, by combining three compliance groups, TSLS estimands will be less affected by weak instrument issues than the LACEs or LAIE because TSLS draws on a larger group of units.

In both of the experiments I analyze here, there is one-sided noncompliance because respondents are never contacted when randomized to receive no contact. Under this design there are no always-takers, and so the coefficients on the lower order terms will be equal to the aggregated LACEs, $\delta_1 = \tau_{c-,1}(0)$ and $\delta_2 = \tau_{-c,2}(0)$. With or without one-sided noncompliance, it is not possible to take the sum of coefficients to estimate different causal contrasts as is possible with a typical linear model. For example, the LAJE is not equal to $\delta_1 + \delta_2 + \delta_3$ because of treatment effect heterogeneity between compliance groups. As is clear from Theorem 2, each coefficient mixes over effects from different joint compliance types, so that their sum would not reflect an effect for the joint compliers. Overall, interpretation of TSLS requires care in these experiments and may not accurately reflect the LACEs and the LAJE. It is important to note that Theorem 2 relies on independence of the instruments, which is not always satisfied and is not required for the identification of the LACEs.

4 Data Analysis of Two GOTV Experiments

4.1 New Haven 1998 GOTV Experiment

In this section, I use the above methods to analyze the New Haven experiment, focusing on one arm of the study that was a 2×2 factorial design as described above. Another arm of the experiment sent mailers to households, but I omit this arm since there is no measurement of compliance with the mailers treatment. Because the phone treatment was randomized differently depending on the mailer treatment, I subset to households that were randomized to the control condition for the mailers arm, leaving 11,689 households. Randomization was done at the household level and so the outcome in this case is the number of voters in the household that turned out in the 1998 general election. Since the sampled households had at most two residents, the treatment effects can range from -2 to 2. Imai (2005) and Gerber and Green (2005) provide a robust discussion of various methodological and technical issues with this particular experiment, most of which are irrelevant to the current discussion, while Hansen and Bowers (2009) analyze this experiment with randomization-based inference.

The New Haven experiment is a good example of a setting where the above assumptions are met, but that the design and compliance rates of the study make it difficult to draw conclusions. Random assignment of the two forms of contact were carried out independently, with $\Pr[Z_{i1} = 1] = 0.1$ and $\Pr[Z_{i2} = 1] = 0.2$ so that the only 5% of the households were randomly selected to receive both forms of contact. Furthermore, I estimate the compliance probabilities in this setting are $\hat{\rho}_{cc} = 0.117$, $\hat{\rho}_{cn} = 0.217$, and $\hat{\rho}_{nc} = 0.127$, meaning that there are very few joint compliers in this example. Thus, while Assumption 1(iv) is satisfied, the effect of Z_{ij} on D_{ij} is very weak in this experiment. Monotonicity holds here by design since those in the control groups are unable to receive contact and so there are no defiers or always-takers. Finally, as discussed above, treatment exclusion appears plausible given the design, but could be violated if randomized assignment into in-person (phone) contact affected whether phone (in-person) contact was actually made. Because of the randomization, this can be tested. In this case, the effect of in-person randomization on phone contact is -0.013 (95% confidence interval: $[-0.073, 0.048]$) and the effect of phone randomization on in-person contact is 0.042 (95% CI: $[-0.022, 0.106]$). Thus, in this particular context there appears to be little evidence of a violation of the treatment exclusion.

Figure 1 shows the estimates of the LACEs, the LAJE, and the LAIE using the plug-in estimators proposed in Section 3 with confidence intervals using the proposed variance estimators from the Supplemental Materials. For both treatments, the Figure additionally shows the overall LATE and the effect estimated from the iTSLS approach of Section 3.1. In the original analysis, Gerber and Green (2000) used a standard IV ratio estimator for each instrument-treatment pair separately and I find estimated LATEs that are very similar to theirs—a positive effect of in-person contact and no effect of phone contact. The basic LACE estimates using the above estimators show a great deal of variation across comparisons and across compliance types. For example, the point estimates of the joint-complier effects of the phone and in-person contact alone, $\widehat{\tau}_{cc,1}(0)$ and $\widehat{\tau}_{cc,2}(0)$, suggest that these forms of contact decrease turnout by a large amount. These point estimates are four to five times larger in magnitude than the estimated effects that aggregate across different compliance groups, $\widehat{\tau}_{c-,1}(0)$ and $\widehat{\tau}_{-c,2}(0)$. Furthermore, with in-person contact, the effects for phone never-takers, $\widehat{\tau}_{nc,2}(0)$ is positive with a 95% confidence interval that excludes zero. The same LACE for the joint compliers is strongly negative, but with a wide confidence interval that includes zero. While is it very difficult to make inferences about the LACEs in this setting, there is suggestive evidence that there is strong heterogeneity in treatment effects by compliance type.

What about the relationship between the plug-in and iTSLS estimates? Given the one-sided noncompliance and the results of Theorem 2, it is unsurprising that the aggregated LACEs are almost identical to the lower-order term estimates of the iTSLS model. On the other hand, the point estimates $\widehat{\tau}_{cc,1}(1)$ and $\widehat{\tau}_{cc,2}(1)$, which are the effects of each contact when a household has been contacted by the other method, show stark contrasts with the iTSLS approach. The iTSLS estimates are obtained by adding the lower-order and interaction terms. But as discussed above, this combines effect estimates from different subpopulations and therefore will be biased for the $\tau_{cc,j}(1)$. In addition to differences in the point estimates, the iTSLS confidence intervals are also narrower than the plug-in estimators since they draw on more compliance groups for estimation.

Shifting to combinations of the basic LACEs, the plug-in estimator and iTSLS produce almost identical estimates for the LAIE, which is expected given Theorem 2. While the point estimates are in favor of the synergistic hypothesis, the uncertainty is far too large in this experiment to learn much of anything about the interaction. With joint effects, there are more striking differences between the proposed estimators and the iTSLS approach. The plug-in LAJE estimate is strongly negative, -0.613, though with high uncertainty, while the joint effect from the iTSLS model finds a joint effect of 0.563 and is statistically significant at the typical levels. Thus, the iTSLS and the LAJE have similar magnitudes but in the opposite direction. Once again, though, the uncertainty is far lower using iTSLS. Overall, the results of the New Haven experiment are muddled by massive uncertainty caused by the experimental design and the low rates of compliance.

One concern that we might have in this setting is that the low rates of compliance are driving significant biases or causing our confidence intervals to have poor performance. To assess this possibility, I simulated data from a one-sided noncompliance design with similar compliance rates, sample size, and assignment probabilities to this experiment. In these simulations, presented in Supplemental Materials, I find that the plug-in estimators to have a bias of roughly 12% of the true parameter value but also have confidence intervals with close to nominal coverage or perhaps slight overcoverage. Thus, while the point estimates here may be suspect due to bias from the weak instrument problem, the confidence intervals should accurately reflect the uncertainty due to low compliance.

4.2 2004 Youth Vote Coalition Experiment

To overcome the issues of low compliance rates in the last experiment, I now turn to an even larger field experiment. Nickerson (2007) reported the results of a multi-site experiment conducted through the Youth Vote Coalition (YVC), a nonpartisan organization dedicated to GOTV efforts for citizens under the age of 30. In each of the 17 counties that served as experimental sites, individual registered voters were randomly assigned in a 2×2 design to receive a call from a volunteer phone bank (Z_{i1}), a professional phone bank (Z_{i2}), both, or neither. Previous research had shown that volunteer phone banks were effective at voter mobilization, but that professional phone banks (like those used in the New Haven experiment) had less efficacy. The outcome of this study was turnout in the 2004 presidential election, as measured by the public voter file. The YVC experiment was designed to compare these two types of calls in the same populations at the same time. For similar reasons to the New Haven experiment, there was noncompliance on calls from both types of phone banks. In the original analysis, Nickerson (2007) used TSLS with no interaction to estimate the LATE for each type of contact separately, but did not consider the potential interaction between the two types of calls. This interaction is important as it could tell us whether multiple forms of GOTV contact are synergistic or whether there are diminishing returns to additional instances of contact.

To explore these issues, I reanalyze this experiment and focus on college-age respondents (ages 17–22 at the time of the call), who are often the subject of these interventions. As with the New Haven experiment, many of the assumptions here are satisfied by the design of the experiment. In analyzing the data, I found treatment exclusion violations in 3 of the experimental sites and 3 that had effectively assigned only one form of contact. I excluded one other site because it had a joint compliance rate of just 2%, though this has little effect on the estimates below. This left 26,974 respondents spread over 10 experimental sites. In a test of treatment exclusion in this group, the effect of volunteer-bank randomization on professional-bank contact is 0.001 (95% confidence interval: $[-0.015, 0.017]$) and the effect of professional-bank randomization on volunteer-bank contact is -0.006 (95% CI: $[-0.021, 0.009]$). Thus, in the remaining experimental sites, there appears to be little evidence of a violation of the treatment exclusion.

Two additional features make the YVC experiment appealing compared to the New Haven experiment. First, each type of phone bank was randomized with probability 0.5 so that each of the four possible assignment vectors had equal probability of occurring. Second, the compliance types are more evenly balanced in this setting, with $\hat{\rho}_{cc} = 0.23$, $\hat{\rho}_{cn} = 0.29$, $\hat{\rho}_{nc} = 0.15$, and $\hat{\rho}_{nn} = 0.33$. Thus, the probability of joint compliance in this experiment is double that of the New Haven experiment. Both of these features should lead to much lower uncertainty compared to the previous experiment. Figure 2 shows the results of this experiment for the joint compliers. Because of the stratified random sampling by site, these results are the averages of the site-specific LACE and LAIE plug-in estimators, weighted by the sample size of site. The confidence intervals are based on the standard combined variance calculation from a stratified randomized experiment (see, for example, Imbens and Rubin, 2015, p. 204), using the variance estimator from the Supplemental Materials for the within-site variance estimates.

It is possible to learn much more about the joint effectiveness of GOTV efforts in this experiment. The results show that both the volunteer and professional phone bank calls had a large positive impact on turnout among respondents, at least when they only received one call ($\tau_{cc,1}(0)$ and $\tau_{cc,2}(0)$). The a single professional phone call, for instance, had a 20 percentage-point effect on turnout. The effects of one phone-bank contact when paired with the other phone bank ($\tau_{cc,1}(1)$ and $\tau_{cc,2}(1)$), on the other hand, were estimated to be negative but not statistically significant. And, in fact, there is a large, negative interaction between these effects as indicated by the LAIE. Thus, there appear to be rather severe diminishing returns to additional instances of GOTV contact in this study, a finding that is in support of the diminishing returns hypothesis and in contrast with much of the previous literature either which has not investigated interactions such as these or have found very small or null ITT interactions. This novel result tells us that in settings like this, it is much more valuable to reach a larger group of individuals rather than concentrate multiple phone calls on a smaller group. It is not clear if these diminishing returns results are due to the two forms of contact being similar (both phone calls) and whether we would see diminishing returns in a properly powered study of other forms of contact.

Finally, these results can be used by campaigns to evaluate the cost-effectiveness of second contact attempts. For instance, the LACE of volunteer contact among noncompliers for professional contact is negative and very close to zero ($\widehat{\tau}_{cn,1}(0) = -0.007$, 95% CI: $[-0.053, 0.039]$), whereas the same effect for joint compliers is large and positive as shown in Figure 2. Thus, if a campaign has made a professional phone bank call and failed to reach a respondent, they can predict that a volunteer call will produce very little effect on turnout. Thus, in future GOTV efforts, the results of these analyses could provide valuable information for campaigns attempting to maximize voter outreach.

5 Conclusion

Causal interaction between different forms of GOTV contact has been understudied in part due to a lack of statistical tools to do so. This paper extends the single-treatment IV framework to allow for the estimation of conditional and interaction effects in factorial experiments with noncompliance. With this framework in hand, the paper was able to provide evidence for the diminishing returns hypothesis in GOTV contact, which has large ramifications for how campaigns should attempt to maximize voter turnout. In another experiment, inference is much more difficult due to the weak instrument problem and a suboptimal experimental design.

The statistical framework here is applicable far beyond GOTV studies—it can be applied to any 2×2 factorial design with noncompliance on each factor. The assumptions needed are a combination of the single-treatment IV assumptions and a treatment exclusion restriction, which helps with identification by limiting the number of principal strata. This paper provides plug-in estimators for the LACEs, which are average comparisons between levels of treatment conditional on compliance types. I show that TSLS can be used to estimate the local average interaction effect and, in experiments with one-sided noncompliance, can also be used to estimate aggregated version of the LACEs.

There are several ways in which this statistical framework could be extended. Most obviously, it would be straightforward to allow for $K > 2$ factors in the framework, though the statistical power to estimate LACEs and LAIEs in that setting may be low. It is also important to develop a theoretical understanding of how to bound treatment effects when the treatment exclusion restriction is violated. Another obvious direction for future research is to develop a similar set of results and estimators for situations with non-binary treatments and instruments and with continuous covariates. For instance, one could extend the local average response function approach of Abadie (2003) to a setting with multiple treatments and multiple instruments.

References

Abadie, Alberto. 2003. "Semiparametric instrumental variable estimation of treatment response models." *Journal of Econometrics* 113(2):231–263.

Angrist, Joshua D. and Guido W. Imbens. 1995. "Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity." *Journal of the American Statistical Association* 90(430):431–442.

Angrist, Joshua D., Guido W. Imbens and Donald B. Rubin. 1996. "Identification of Causal Effects Using Instrumental Variables (with discussion)." *Journal of the American Statistical Association* 91:444–455.

Arceneaux, Kevin and David W. Nickerson. 2009. "Who Is Mobilized to Vote? A Re-Analysis of 11 Field Experiments." *American Journal of Political Science* 53(1):1–16.

Bound, John, David A Jaeger and Regina M Baker. 1995. "Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogeneous Explanatory Variable is Weak." *Journal of the American Statistical Association* 90(430):443.

Cheng, Jing and Dylan S Small. 2006. "Bounds on causal effects in three-arm trials with non-compliance." *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 68(5):815–836.

Clarke, Paul S. and Frank Windmeijer. 2010. "Identification of causal effects on binary outcomes using structural mean models." *Biostatistics* 11(4):756–770.

Fowler, Anthony. 2013. "Electoral and policy consequences of voter turnout: Evidence from compulsory voting in Australia." *Quarterly Journal of Political Science* 8(2):159–182.

Frangakis, Constantine E. and Donald B. Rubin. 2002. "Principal stratification in causal inference." *Biometrics* 58:21–29.

Gerber, Alan S. and Donald P. Green. 2000. “The effects of canvassing, telephone calls, and direct mail on voter turnout: A field experiment.” *American Political Science Review* 94(03):653–663.

Gerber, Alan S. and Donald P. Green. 2005. “Correction to Gerber and Green (2000), replication of disputed findings, and reply to Imai (2005).” *American Political Science Review* 99(02):301–313.

Gerber, Alan S., Donald P. Green and Christopher W. Larimer. 2008. “Social Pressure and Voter Turnout: Evidence from a Large-Scale Field Experiment.” *The American Political Science Review* 102(1):33–48.

Green, Donald P., Alan S. Gerber and David W. Nickerson. 2003. “Getting Out the Vote in Local Elections: Results from Six Door-to-Door Canvassing Experiments.” *Journal of Politics* 65(4):1083–1096.

Green, Donald P., Mary C McGrath and Peter M Aronow. 2013. “Field Experiments and the Study of Voter Turnout.” *Journal of Elections, Public Opinion & Parties* 23(1):27–48.

Hansen, Ben B and Jake Bowers. 2009. “Attributing effects to a cluster-randomized get-out-the-vote campaign.” *Journal of the American Statistical Association* 104(487):873–885.

Hernán, Miguel A. and James M. Robins. 2006. “Instruments for causal inference: an epidemiologist’s dream?” *Epidemiology* 17(4):360–372.

Imai, Kosuke. 2005. “Do get-out-the-vote calls reduce turnout? The importance of statistical methods for field experiments.” *American Political Science Review* 99(02):283–300.

Imbens, Guido W. and Donald B. Rubin. 2015. *Causal Inference for Statistics, Social, and Behavioral Sciences*. Cambridge University Press.

Imbens, Guido W. and Joshua D. Angrist. 1994. “Identification and Estimation of Local Average Treatment Effects.” *Econometrica* 62(2):467.

Nickerson, David W. 2007. "Quality Is Job One: Professional and Volunteer Voter Mobilization Calls." *American Journal of Political Science* 51(2):269–282.

Obama for America. 2013. 2012 Obama Campaign Legacy Report. Technical report. **URL:** *http://secure.assets.bostatic.com/frontend/projects/legacy/legacy-report.pdf*

Robins, James M. 1994. "Correcting for non-compliance in randomized trials using structural nested mean models." *Communications in Statistics* 23(8):2379–2412.

Rubin, Donald B. 1974. "Estimating causal effects of treatments in randomized and nonrandomized studies." *Journal of educational Psychology* 66(5):688.

VanderWeele, Tyler. 2015. *Explanation in causal inference: methods for mediation and interaction*. Oxford University Press.

C_i	$D_{i1}(0)$	$D_{i1}(1)$	$D_{i2}(0)$	$D_{i2}(1)$	LACE(s) for this strata
cc	0	1	0	1	$\tau_{cc,1}(0) = E[Y_i(1,0) - Y_i(0,0) \mid C_i = cc]$ $\tau_{cc,1}(1) = E[Y_i(1,1) - Y_i(0,1) \mid C_i = cc]$ $\tau_{cc,2}(0) = E[Y_i(0,1) - Y_i(0,0) \mid C_i = cc]$ $\tau_{cc,2}(1) = E[Y_i(1,1) - Y_i(1,0) \mid C_i = cc]$
cn	0	1	0	0	$\tau_{cn,1}(0) = E[Y_i(1,0) - Y_i(0,0) \mid C_i = cn]$
nc	0	0	0	1	$\tau_{nc,2}(0) = E[Y_i(0,1) - Y_i(0,0) \mid C_i = nc]$
ca	0	1	1	1	$\tau_{ca,1}(1) = E[Y_i(1,1) - Y_i(0,1) \mid C_i = ca]$
ac	1	1	0	1	$\tau_{ac,2}(1) = E[Y_i(1,1) - Y_i(1,0) \mid C_i = ac]$
aa	1	1	1	1	None
an	1	1	0	0	None
na	0	0	1	1	None
nn	0	0	0	0	None

Table 1: Joint compliance types (principal strata) under treatment exclusion and monotonicity and the local average conditional effects (LACEs) associated with each.

$f_{11\|11} =$	$\rho_{cc} + \rho_{aa} + \rho_{ac} + \rho_{ca}$	$f_{11\|10} =$	$\rho_{aa} + \rho_{ca}$
$f_{10\|11} =$	$\rho_{cn} + \rho_{an}$	$f_{10\|10} =$	$\rho_{cc} + \rho_{an} + \rho_{ac} + \rho_{cn}$
$f_{01\|11} =$	$\rho_{nc} + \rho_{na}$	$f_{01\|10} =$	ρ_{na}
$f_{00\|11} =$	ρ_{nn}	$f_{00\|10} =$	$\rho_{nc} + \rho_{nn}$
$f_{11\|01} =$	$\rho_{aa} + \rho_{ac}$	$f_{11\|00} =$	ρ_{aa}
$f_{10\|01} =$	ρ_{an}	$f_{10\|00} =$	$\rho_{an} + \rho_{ac}$
$f_{01\|01} =$	$\rho_{cc} + \rho_{nc} + \rho_{na} + \rho_{ca}$	$f_{01\|00} =$	$\rho_{na} + \rho_{ca}$
$f_{00\|01} =$	$\rho_{cn} + \rho_{nn}$	$f_{00\|00} =$	$\rho_{cc} + \rho_{cn} + \rho_{nc} + \rho_{nn}$

Table 2: Relationship between observed strata probabilities, $f_{d_1 d_2|z_1 z_2} = \Pr[D_i = (d_1, d_2)|Z_i = (z_1, z_2)]$, and compliance types probabilities, $\rho_{km} = \Pr[C_i = km]$, under Assumptions 1 and 2.

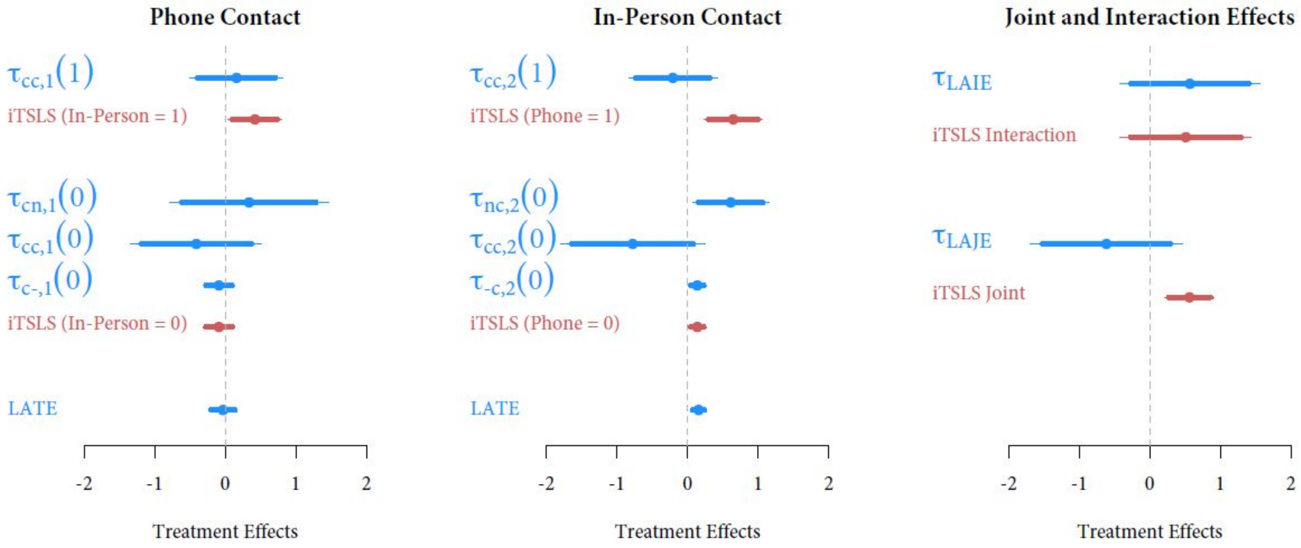

Figure 1: Estimation of LACEs of phone contact (left) and in-person contact (middle), along with the LAJE and LAIE (right). Thin (thick) lines are 95% (90%) confidence intervals, respectively, using the proposed variance estimators proposed. Estimates using an interacted TSLS model and the overall LATE for each treatment are in red. $N = 11,689$ households.

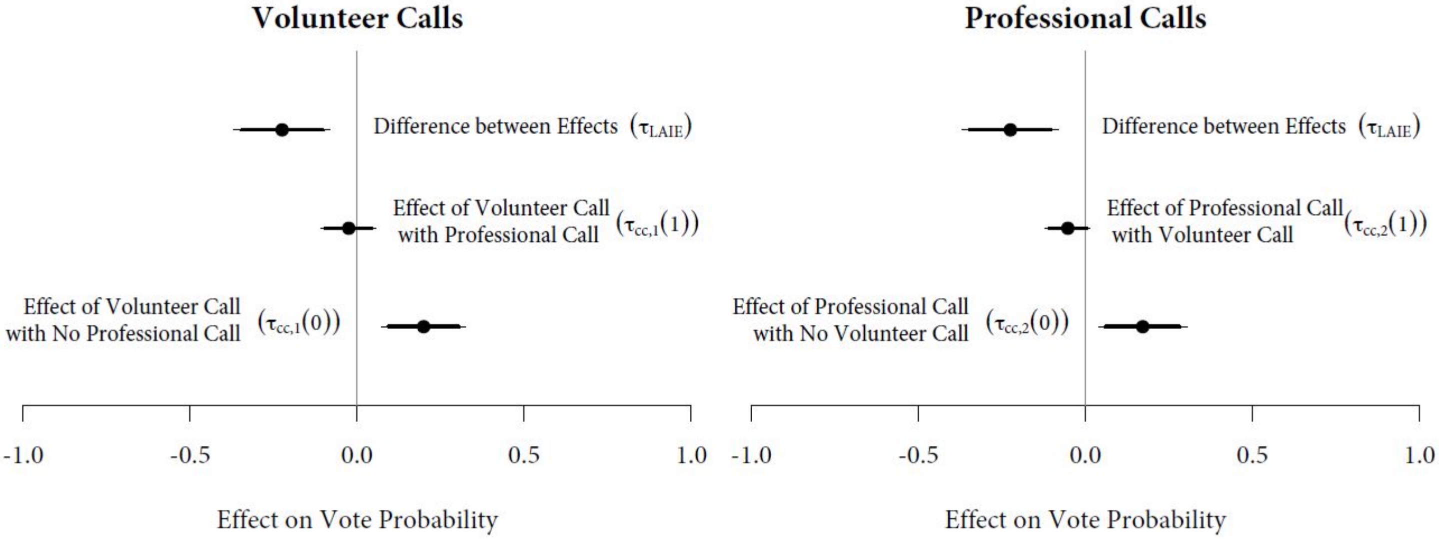

Figure 2: Combined LACE and LAIE estimates among joint compliers across experimental sites from the Nickerson (2007) youth GOTV experiment. Thin (thick) lines are 95% (90%) confidence intervals based on pooled variance across the sites. $N = 26,974$ registered voters aged 17–22.